Firm Specialization vs. Diversification*

Anmol Bhandari[1], Sanhitha Jugulum[2], Laura Veldkamp[3], Venky Venkateswaran[4]

[1]University of Minnesota and NBER
[2]Columbia Business School
[3]Columbia Business School and NBER
[4]NYU Stern and NBER

November 2, 2026

Abstract

This paper investigates the fundamental trade-off between firm specialization and diversification. While management theories emphasize core competency to maximize efficiency, finance favors diversification to mitigate cash flow volatility. We develop a heterogeneous-firm general equilibrium model to quantify these forces. We devise a novel empirical approach, grounded in theory, to measuring them. We first extract granular revenue shares from 10-K filings, using a large language model. Next, we use the co-movement between sectoral aggregates as (an inverse) measure of the economic distance between them. The observed relationship between scope and markups at the firm-level pins down the price of risk. These sufficient statistics also recover productivity at the firm-sector level. The results reveal that risk considerations lower aggregate output significantly (by 10%). We also find that firms are more likely to operate in related sectors, where they tend to have higher productivity on average.

*We thank Jaedo Choi (discussant), seminar participants at the Federal Reserve Bank of St. Louis, SED 2026, the Columbia Macro Lunch and the Texas Monetary for helpful comments.

1 Introduction

Should a firm deepen its expertise by producing related products, or diversify into new and distinct lines? The fields that study this choice most directly offer guidance, but they disagree. A key recommendation from management science is that firms focus on their core competency, specializing in related activities to exploit shared technologies, brands, and know-how (Rumelt, 1974; Prahalad and Hamel, 1990). This view elevates Adam Smith's idea that specialization is the source of efficiency, to the level of the firm. Finance, in contrast, makes the case for diversification. In a frictionless world the firm itself would have no reason to manage risk, since its shareholders can diversify on their own account. But costly external finance, bankruptcy costs, and owners and managers whose wealth is concentrated in the firm leave firms bearing risk that markets do not absorb; expanding into different businesses is then a way to reduce it (Lewellen, 1971; Stein, 1997; Maksimovic and Phillips, 2002). Both arguments are compelling, yet they push in opposite directions. The firm dynamics literature in macroeconomics and finance, where the aggregate consequences of these choices would matter, is largely silent: in workhorse models such as Klette and Kortum (2004) and their many descendants, firms grow by innovating and introducing new goods, driving creative destruction and aggregate productivity, but the goods themselves arrive at random, so the models take no stand on which products a firm chooses to produce. How firms choose to specialize or diversify, and the aggregate consequence of those decisions, remains an open question.

In this paper, we develop a heterogeneous firm general equilibrium model to analyze and quantify these forces. The model uses standard components from macro and finance, and guides our development of novel empirical measures of risk and gains to specialization at the firm-level. This allows us to quantitatively assess the micro and macro implications of this trade-off. We find that product choice and risk considerations have significant aggregate consequences. In particular, the compensation firms demand for bearing undiversified risk acts as a tax on production, one that falls most heavily on the most productive firms, whose expertise is concentrated in correlated activities. This lowers output by 10% relative to an otherwise identical economy without risk. These results offer a new perspective on the cost of risk: Risk is costly for economies because it induces firms to take risk-reduction measures, like broadening scope into less related, less productive activities, that compromise output.

In the model, firms choose the amount of labor input for each of the sectors in their portfolio

before observing shocks to sectoral demand and productivity. These shocks are correlated, so from the firms' perspective, sectors differ both in their productivity as well as riskiness. Producing in sectors that are similar to a firm's home sector leverages its existing expertise, resulting in higher productivity. But, profits from such activities are also riskier because they are highly correlated with firm's overall cashflow risk. Sectors that are less similar to the firm's core activities, on the other hand, reduce exposure to shocks. The optimal decision can be interpreted in terms of expected markups: firms demand compensation for risk by charging a higher gap between prices and conventional production marginal costs. Thus, the theory generates endogenous variation in markups by firm and sector (even when the demand curve exhibits a constant elasticity of substitution). The firm's optimal markup in a given sector is a function of the sector's *risk exposure* and the *price of risk*. The former depends on the sectoral composition of the firm's revenue streams: specifically, the revenue-share weighted average correlation of the sector with the rest of the firm's portfolio. The higher the average correlation, the larger the sector's risk exposure for the firm and therefore, the higher the optimal markup.

When aggregated to the firm-level, this produces a tight connection between the firm-level markup and its *scope*, the model-implied measure of diversification. Scope is defined as 1 minus the weighted average risk-exposure of the various sectors the firm operates in. This offers a new interpretation of the breadth of a firm's operations, by going beyond a simple count and explicitly accounting for how different they are from each other. Concretely, the theory points to a measure of scope that equals the expected distance between two randomly drawn dollars of the firm's revenue, where the distance between two sectors, $d_{jk} = 1 - \rho_{jk}$, measures how weakly their revenues comove. A broader scope implies lower risk through diversification and is therefore associated with lower markups in equilibrium.

These predictions follow from firms' productivity–risk trade-off, offering a novel explanation for markup and scope patterns across firms and industries. It is worth highlighting the difference between scope and firm size. Larger firms typically charge *higher* markups, reflecting scale economies or market power. In contrast, firms with greater scope, that is, those that produce a broader and more diversified set of goods, charge *lower* markups, consistent with diversification lowering the cost of risk-bearing; this negative scope–markup relationship survives allowing markups to load directly on market share. This distinction uncovers a new margin of heterogeneity in markups and risk exposure.

To quantify the output lost due to risk, we compare the aggregate output in the calibrated

economy to a counterfactual one with the same productivities but a price of risk equal to zero. The model yields an intuitive, tractable expression for this object, as a function of the joint distribution of productivities and risk wedges in equilibrium. Recovering this distribution requires the measurement of three objects: 1) firm scope, 2) a price of risk based on firms' revealed preference, and 3) firm-sector productivities.

To construct firm scope, we need (i) each firm's revenue shares across sectors and (ii) correlation matrix of sectoral shocks. The former cannot be easily computed with existing approaches or data. For example, in their seminal paper, Hoberg and Phillips (2025) significantly advanced the measurement of firm scope using textual analysis. Their approach yields the number of sectors a firm operates in, but cannot be mapped to within-firm revenue shares. Our strategy, instead, uses the power of frontier large-language models (LLMs) to directly infer revenue shares from the firm's 10-K filings. We implement a two-stage procedure: the first stage distills the business, risk-factor, and management's discussion sections of each filing into a structured summary of the firm's revenue-generating activities. We store these summaries, so that the analysis can be replicated even as the underlying models change. The second stage reads a firm's summaries from filings five years apart and returns the set of active 4-digit NAICS sectors, with revenue shares in each year. We validate the output in several ways. In the subsample of filings whose narrative contains no quantitative segment information, revenue shares inferred from the text alone closely match those constructed with reported segment data as an anchor: the two agree on the firm's top 4-digit sector in 76% of cases (87% at the 2-digit level), and confirmed misclassifications amount to under 2% of the sample. This novel dataset has applicability beyond our question of interest here. [1]

The second ingredient – the correlation structure of sectoral shocks – is more straightforward to estimate using sectoral aggregate data. The use of a relatedness measure is a key novel feature of our approach and distinguishes us from other approaches for assessing the breadth of a firm's activities. It is also central to our goal of assessing diversification. Simply measuring the number of sectors a firm operates in, as some existing measures of scope do, is obviously not sufficient: it would be like judging portfolio risk based on the number of stocks held. Just as the covariance properties of the stocks are an essential ingredient in that case, our risk-based notion of scope requires the use of a risk-based measure of distance. Combining these two inputs delivers the scope

[1] For example, it can be used to construct more robust measures of firm-level exposure to sectoral shocks. It can also serve as the basis for projecting parameter estimates at the sector level – such as markups, nominal rigidity, innovation spillovers or input-output linkages – to the firm-level.

measure directly. As an external check, we show that scope predicts survival: a one-standard-deviation increase in $\ln(1 + Scope)$ is associated with a roughly 5% lower probability of exiting the sample over the following five to ten years.

Next, to measure the price of risk, we make use of a key prediction of the theory: the relationship between scope and markups is moderated by the firm's price of risk. Specifically, a higher risk price makes markups more sensitive to diversification and therefore, to scope. Based on this insight, we use the observed (within-firm) variation in scope and markups to estimate the price of risk through a generalized method of moments procedure; instrumenting scope with its five-year lag, which purges contemporaneous shocks that move measured markups and measured scope together, delivers nearly identical estimates. This is a novel approach: while risk price measures abound, most are derived from equity markets, reflecting the preferences and constraints of equity traders. Our measure reflects the revealed preference of firms trading off their own return for risk reduction through scope expansion.[2] The results imply that, for the average firm in our sample, the *risk wedge* scales the benchmark (i.e. riskless) markup by a factor of 1.14.

We then use the model equations to recover productivities at the firm-sector level from market shares. In addition to being a key input in the aggregate loss due to risk, this detailed productivity profile allows us to quantify the gains to specialization, i.e. how much more productive firms are in activities close to their expertise. To do this, we adopt a low-dimensional representation of productivity: the productivity of a firm in a given sector is a function of its "home-sector" productivity and the distance of the sector from the home sector. The rate of decay, i.e. how quickly a firm's productivity declines as it moves away from its home sector, thus can be mapped directly to gains from specialization. Our estimates suggest that these gains are significant: productivity drops by roughly 60 log points on stepping outside the home sector, and by a further 15 log points between the closest and the furthest sectors. This representation is both interpretable and portable: our formulation and estimates can be embedded easily in firm dynamics models, making this novel empirical contribution of potentially broader interest to the firm dynamics literature. The data also line up with these estimates on the extensive margin: a firm is about 2.2 times more likely to operate in a sector highly correlated with its home sector (correlation 0.9) than in a distant one (correlation -0.9), direct evidence against the random product arrival assumed in workhorse firm-dynamics models.

[2] While firms' choices maximize equity values in ideal settings, incentive misalignment, financing constraints, and other frictions can break this link, causing the firm's internal price of risk to differ from the equity holders'.

Finally, Section 5 combines these ingredients to estimate the aggregate cost of risk considerations in firm input choices. We estimate that, relative to a setting with a zero risk price, aggregate output is 10% lower. To gain further intuition, we decompose this total loss into a component attributable to the average wedge and one that stems from the correlation between risk and productivity. When this correlation is positive – i.e. the most productive firms demand the most compensation for risk – the output losses from risk are amplified (relative to a counterfactual economy in which the risk wedge and productivity are un-correlated). This pattern is driven partly by the structure of firm-sector productivity discussed above: when gains to specialization are large, high productivity firms in a sector tend to be less diversified and therefore, require higher risk compensation. The data point to a large role for this amplification channel – about half of the total loss from risk can be attributed to the covariance between risk and productivity.

1.1 Contribution to the existing literature

Our paper unifies three historically distinct perspectives on firm scope through a measurable construct of product distance and revenue covariance. By doing so, we bridge the gap between the finance view of diversification, the management view of core competencies, and the macro-IO view of multi-product dynamics.

In corporate finance, the study of firm scope is primarily framed through the lens of risk management and internal capital markets. Early work by Lewellen (1971) and more recently Stein (1997) suggests that firms expand across activities to mitigate financing frictions and smooth idiosyncratic shocks. Maksimovic and Phillips (2002) and Matvos and Seru (2014) further explore how diversified firms allocate resources across segments, often finding that internal capital markets can provide stability even if they lead to a valuation discount. Closely related, Boar et al. (forthcoming) quantify the aggregate output losses from uninsurable idiosyncratic business risk and find them large, at 15.4%. Our roughly 10% loss is comparable in magnitude but isolates a distinct channel: the productivity cost of the scope expansion that firms undertake to diversify that risk. We complement this work with a production economy that also features gains to specialization, estimating how the two forces trade off.

A growing literature reinterprets objects usually read as market power or misallocation as compensation for risk. David et al. (2022) show that firm-specific risk premia generate dispersion in the marginal product of capital that mimics misallocation but is an efficient adjustment for risk, and that this risk-based dispersion lowers long-run aggregate productivity; relatedly, David

et al. (2016) show that imperfect information about productivity produces persistent dispersion in marginal products. Di Tella et al. (2025) formalize the markup counterpart: when production is risky and risk sharing is incomplete, the wedge between price and marginal cost is a "risk markup" rather than a market-power markup, with distinct implications for optimal policy. This complements the evidence in Boar and Midrigan (2023) that much of cross-sectional markup dispersion reflects compensation for risk. We share this lens, but add a product-space dimension. In our model the markup contains a risk wedge that compensates the firm for bearing undiversified revenue risk, and this wedge is a function of the firm's *scope*, so the cross-section of markups maps to the cross-section of diversification, and the aggregate cost we measure operates through the productivity penalty of risk-driven scope expansion. Our structural counterfactual thereby also speaks to the macro literature on the level and dispersion of markups (Edmond et al., 2023): it provides a new decomposition of measured markups, identifying how much of their dispersion is driven by heterogeneity in firms' scope across the product space.

In contrast to the finance view, management and organizational scholars emphasize the costs of straying too far from a firm's core. Rumelt (1974) and the seminal work by Prahalad and Hamel (1990) argue that firms should focus on their core competencies to maintain organizational coherence and efficiency. Teece et al. (1994) formalized the notion that firms are constrained by their existing knowledge and routines, a finding echoed by empirical evidence from Bresnahan et al. (2011), who show that firms expand most successfully into technologically or organizationally proximate areas. Our model translates these qualitative notions of "coherence" and "relatedness" into a quantitative distance parameter (θ), allowing us to quantify the trade-off between the gains from specialization and the gains from diversification.

The macroeconomics and industrial organization literatures typically treat scope as the number of products produced or sectors served, sometimes linked through firm-level constraints. Foundational models such as Bernard et al. (2010) and Mayer et al. (2016) focus on product switching and how firms sort across products based on productivity. However, many models follow a Klette and Kortum (2004) or Eaton and Kortum (2002) structure where products arrive randomly, making the models silent on the composition of firms' products. We extend this literature by introducing an explicit product space and showing why the distance of firms' products from each other matters for aggregate output.

Our model builds on the multi-attribute product space pioneered by Pellegrino (2025), to explain various industrial organization pricing and elasticity facts. Our key departure is the intro-

duction of firms that price revenue risk. This is essential for studying the costs and benefits of product diversification.

Our ideas are most closely related to recent theories of innovation. Berlingieri et al. (2025) characterizes innovation as happening in "bursts," where a single discovery leads a firm to introduce multiple new products at once, to capture market power. Furthermore, De Ridder (2021) argues that information technology increases the fixed costs of industry entry while decreasing marginal costs, which incentivizes firms to specialize and narrow their production scope. Ma et al. (2024) show that business concentration has risen, in a way consistent with the growing importance of intangible assets. Our approach estimates these forces as gains to specialization and complements this work by examining how these types of gains compete with diversification.

Methodologically, our work contributes to the burgeoning field of text-based analysis in economics. We build on the pioneering work of Hoberg and Phillips (2016), who used textual similarity in 10-K filings to define product differentiation and competitors. However, whereas Hoberg and Phillips (2025) and Kim et al. (2023) often rely on cosine similarity or transformer-based embeddings to measure linguistic proximity, we innovate in four ways: (1) we map firms to NAICS sectors, to better align with macroeconomic data. (2) More conceptually, we extract firm-sector revenue shares. This allows us to describe how much of each sector's risk a firm is exposed to. (3) We estimate firm-sector productivity, which is an essential input into our estimate of the cost of firm uncertainty and could be useful for other applications. (4) Using revenue covariance, we measure the distance of sectors from each other. Instead of just counting up the number of sector codes assigned to the firm's establishments, we distinguish between operating in nearby industries, or branching out into business lines that are less related.

Finally, Autor et al. (2020) document a recent decline in measured firm scope at the 4-digit industry level; our distance-based measure provides a richer perspective on these trends, emphasizing that reductions in nominal scope need not imply lower diversification when products move within nearby industries. Relatedly, Ekerdt and Wu (2024) document a four-decade reallocation of U.S. manufacturing production toward specialized firms, a key driver of rising industry concentration, and explain it with a demand-side theory of quality in which each firm's scope is taken as given, whereas we model the scope choice itself and quantify its aggregate cost. Their measure, the effective number of industries (the reciprocal of the Herfindahl index of a firm's shares across 6-digit NAICS industries in Census establishment records), treats all industry pairs symmetrically, while ours weights each pair by the comovement of its shocks and is built from consolidated rev-

enue shares recovered from the 10-K rather than from domestic establishment data.

Eberly et al. (2023) define scope in terms of input applicability across tasks, while our framework focuses on outputs and revenue. One could also study how firms diversify input risk and the costs of doing so, with a similar methodology.

Measuring distance has deep roots in the patent literature: Akcigit et al. (2013) and Cai and Li (2019) use patent linkages to trace technological proximity.[3] We extend these insights from patents to products, capturing market-based rather than invention-based proximity. This distinction is important: Most patents are never brought to market and most products are not patented.

2 A Model of Firm Scope

Our environment is a canonical heterogeneous-firm equilibrium production economy, with two key departures: 1) the existence of a product space with correlated shocks, and 2) the assumption that firms care about profit variability, rather than just expected profits. Both are essential to connect to diversification motives. Both are also realistic. Risk management practices are central to the way corporate decision makers are trained.

2.1 Model setup

The economy has a unit mass of *ex-ante* symmetric sectors, indexed by j. Within each sector, a unit mass of firms, indexed by i, produces differentiated varieties. Sectoral output is a composite of different varieties, using a CES aggregator:

$$Q_j = \left(\int Q_{ij}^{\frac{\eta-1}{\eta}} di \right)^{\frac{\eta}{\eta-1}} . \tag{1}$$

The single final good, denoted Q, is a CES-composite of the sectoral output:

$$Q = \left(\int B_j^{\frac{1}{\theta}} Q_j^{\frac{\theta-1}{\theta}} dj \right)^{\frac{\theta}{\theta-1}} . \tag{2}$$

where $B_j > 0$ is a sector-specific preference shock. The shocks are normalized so that $\mathbb{E}[B_j^{\frac{1}{\theta}}] = 1$.

Preferences. The representative household has preferences over consumption and leisure, with a

[3] Related work values these intangible assets (Ayyagari et al., 2023; Eisfeldt and Papanikolaou, 2014) and their macro roles (Acemoglu et al., 2018).

utility function $u(C, L) = \ln C - L$. Normalizing the price of the final good to 1, the household budget constraint is

$$C = WL + \Pi \,,$$

where Π denotes aggregate firm profits.

Technology, inputs, and costs. Production by firm i in sector j is given by $Q_{ij} = A_j A_{ij} L_{ij}$ where A_j denotes a sectoral TFP shock and A_{ij}, the idiosyncratic (i.e. firm-sector) productivity. We set $\mathbb{E}[A_j] = 1$. Labor is traded in a competitive market at a wage W.

Timing of decisions. Each firm chooses labor input $\{L_{ij}\}$ observing its own productivity $\{A_{ij}\}$ but before learning the realization of the sectoral shocks A_j and $\{B_j\}$. The household makes its consumption decisions after all shocks are observed, so realized prices for the sectoral composite and the firm's variety are given by

$$P_j = B_j^{\frac{1}{\theta}} \left(\frac{Q_j}{Q} \right)^{-\frac{1}{\theta}} \,, \tag{3}$$

$$P_{ij} = \left(\frac{Q_{ij}}{Q_j} \right)^{-\frac{1}{\eta}} P_j \,. \tag{4}$$

Equation (4) thus defines the inverse demand for each variety. Combining (4) with the production function, firm i's revenue from its activities in sector j becomes

$$P_{ij} Q_{ij} = (A_{ij} L_{ij})^{1-\frac{1}{\eta}} \underbrace{A_j^{1-\frac{1}{\eta}} B_j^{\frac{1}{\theta}} Q_j^{\frac{1}{\eta}-\frac{1}{\theta}} Q^{\frac{1}{\theta}}}_{\equiv\ X_j} = (A_{ij} L_{ij})^{1-\frac{1}{\eta}} X_j \,. \tag{5}$$

The composite random variable X_j is a summary statistic for the risk related to sector-j faced by the firm. It is an equilibrium revenue shifter and depends on the choices of all firms in the sector and the two exogenous shocks, which shift the demand curve and therefore, the marginal revenue product of labor. We abstract from aggregate uncertainty, so the economy-wide composite Q is deterministic. We define $\bar{X} = \mathbb{E}[X_j]$ and $\sigma^2 = \mathrm{Var}(\ln X_j)$.

Risk and the firm's problem. Firms dislike revenue risk and so have a motive to diversify. We capture this with a standard expected-utility criterion: the firm's decision-maker has a Bernoulli utility u_i (increasing and concave) over consumption $c_i = b_i + \Pi_i$, where b_i is non-stochastic background wealth of the decision maker and $\Pi_i = \sum_j (P_{ij} Q_{ij} - W L_{ij})$ is firm profit, and chooses inputs to maximize $\mathbb{E}[u_i(c_i)]$. The concavity of u_i penalizes payoff dispersion, which is the source of the

diversification motive. In Appendix A we show that, to second order in the size of revenue risk, this problem is equivalent to a mean-variance objective,

$$\max_{\{L_{ij}\}} \quad \mathbb{E}\left[\sum_j (P_{ij}Q_{ij} - WL_{ij})\right] - \gamma_i \operatorname{Var}\left(\sum_j (P_{ij}Q_{ij} - WL_{ij})\right), \tag{6}$$

where the risk-aversion coefficient is one-half the decision-maker's coefficient of absolute risk aversion, evaluated at expected consumption $\bar{c}_i \equiv \mathbb{E}[c_i]$:

$$\gamma_i \;\equiv\; -\tfrac{1}{2}\,\frac{u_i''(\bar{c}_i)}{u_i'(\bar{c}_i)} \;\geq\; 0. \tag{7}$$

The expectation and variance are taken over the sectoral revenue shifters X_j, conditional on the firm's information.[4]

The same appendix shows that a first-order log-linearization of revenue around $X_j = \mathbb{E}[X_j]$ expresses the variance through the covariance of *log* revenue shifters, giving the operational form of the objective:

$$\max_{\{L_{ij}\}} \quad \sum_j (R_{ij} - WL_{ij}) - \gamma_i \sum_j \sum_k R_{ij} R_{ik} \rho_{jk} \sigma^2, \tag{8}$$

where $R_{ij} \equiv \mathbb{E}[P_{ij}Q_{ij}]$ is expected revenue, $\rho_{jk} \equiv \operatorname{Corr}(\ln X_j, \ln X_k)$ is the correlation between the sector-j and sector-k revenue shifters, and $\sigma^2 \equiv \operatorname{Var}(\ln X_j)$.

The coefficient γ_i in (7) or (8) is a local measure of *absolute* risk aversion, carrying units of inverse dollars: the objective subtracts a (dollar-squared) variance from (dollar) profits, so γ_i must convert one into the other. Later we show that γ_i enters the firm's risk-adjusted choices only through the dimensionless product $2\gamma_i R_i$, where $R_i \equiv \sum_j R_{ij}$ is expected revenue; we will denote this product as the firm-level *price of risk*.

Equilibrium: An equilibrium in this economy consists of (i) labor choices $\{L_{ij}\}$ that are optimal for firms, i.e. solve (8), (ii) sectoral and economy-wide quantities, $\{Q_j\}$ and Q respectively, that are consistent with the firm-level choices, i.e. satisfy (1)-(2), and (iii) wages W that clear labor markets, i.e. satisfy the household optimality condition $Q = W$.

[4]This mean-variance objective is consistent with empirical corporate finance evidence on firms' decisions (Eckbo, ed, 2008) and with macro evidence that firms increase investment and sales in response to a randomized data treatment (Kumar et al., 2023). The concavity of u_i could stem from bankruptcy costs, risk-aversion of a manager with equity compensation, or other sources of curvature. Rather than take a stand on a specific driver, we estimate the price of risk directly from the data.

2.2 Equilibrium: Risk, Scope and Markups

It turns out to be instructive to express the solution to the firm's problem in terms of markups, which depend, in addition to the standard substitution elasticities, also on the firm's risk exposure and risk price. This expression will also motivate our firm scope measure and forms the basis for our measurement. The optimal expected markup for firm i in sector j is (derived in Appendix B):

$$\frac{R_{ij}}{WL_{ij}} = \frac{1}{1-\tau_{ij}} \ \frac{\eta}{\eta-1}\,, \tag{9}$$

where the factor $\frac{1}{1-\tau_{ij}}$ is a risk wedge. It scales up the standard markup, $\frac{\eta}{\eta-1}$, and indexes the compensation required by the firm for bearing the risk from its sector j activities. The term τ_{ij} is the product of a (firm-specific) price of risk and (firm-sector-specific) exposure:

$$\tau_{ij} \equiv RiskPr_i \cdot \sigma^2 \cdot RiskExp_{ij} \tag{10}$$

$$\text{where} \qquad RiskPr_i \equiv 2\gamma_i R_i \ , \tag{11}$$

$$RiskExp_{ij} \equiv \sum_k \frac{R_{ik}}{R_i}\rho_{jk}\,. \tag{12}$$

The price of risk increases with the firm's sensitivity to revenue risk, γ_i, and its size (expected total revenue $R_i \equiv \sum_j R_{ij}$). Together with the variance of sectoral risk, σ^2, and the risk exposure $RiskExp_{ij}$, it determines the firm-sector wedge τ_{ij}. The risk exposure is the revenue-weighted average correlation between sector j and the other sectors in the firm's portfolio. It measures how much additional risk sector j contributes to the firm's overall revenue risk. If j is highly correlated with the rest of the firm's portfolio (high ρ_{jk}), then $RiskExp_{ij}$ is high, which drives up τ_{ij} and therefore the optimal markup. Intuitively, hiring more labor in a sector that is highly correlated amplifies firm-level risk, so the firm requires higher markups as compensation. Conversely, sectors that provide diversification benefits reduce effective risk exposure and see lower expected markups.

Summing revenues and labor costs across the firm's sectors—each priced as in (9)—yields the firm-level markup:

$$\frac{R_i}{\sum_j WL_{ij}} = \frac{1}{1 - RiskPr_i \cdot \sigma^2 \cdot (1 - Scope_i)} \ \frac{\eta}{\eta-1}\,, \tag{13}$$

$$\text{where} \qquad Scope_i = 1 - \sum_j \sum_k \frac{R_{ij}}{R_i}\frac{R_{ik}}{R_i}\rho_{jk}\,, \tag{14}$$

measures how diversified firm i's activities are across the product space. It is one minus the revenue-weighted average correlation among the sectors in which the firm operates—each pair (j,k) weighted by the product of its revenue shares $\frac{R_{ij}}{R_i}\frac{R_{ik}}{R_i}$. Scope is high when the firm both spreads its revenue across many sectors *and* operates in sectors whose demand shocks are weakly correlated, so that a downturn in one line is cushioned by the others. A firm concentrated in a single sector—or in several sectors whose shocks move together—has $Scope_i = 0$; a firm whose revenues are spread across sectors with mutually uncorrelated shocks attains the fully diversified benchmark, $Scope_i = 1$. Scope thus captures diversification in the sense that matters for risk: the extent to which a firm's revenue streams hedge one another, not merely how many sectors a firm spans.

Because the revenue weights sum to one, (14) can equivalently be written $\sum_{j,k}\frac{R_{ij}}{R_i}\frac{R_{ik}}{R_i}(1-\rho_{jk})$, which is *Rao's quadratic entropy* (Rao, 1982) with sectoral distance $d_{jk} = 1-\rho_{jk}$: the expected distance between two randomly drawn revenue dollars. This distance interpretation nests several familiar measures. Replacing d_{jk} with the discrete metric ($d_{jk} = 1$ for $j \neq k$, $d_{jj} = 0$) reduces the measure to $1 - \text{HHI}$, the complement of the revenue-based Herfindahl index, while a Hamming-style metric on the hierarchical NAICS structure would instead capture proximity in the classification system itself. Our correlation-based distance is the one the theory singles out as relevant for risk.

Equation (13) shows why this diversification shapes markups. Markups are, in part, compensation for bearing revenue risk, so they rise with the firm's *undiversified* risk, $1 - Scope_i$: the wedge $\frac{1}{1-RiskPr_i\,\sigma^2(1-Scope_i)}$ is decreasing in $Scope_i$. As a firm broadens its scope, it carries less revenue risk and demands a smaller premium in its prices, lowering its markup. Two firms with identical demand elasticities can therefore charge different markups purely because they differ in scope—the more diversified firm charges less. Measured markups thus reflect not only demand conditions but also the riskiness of the firm's portfolio, so markup dispersion can arise from diversification motives even when elasticities are constant.

2.3 Aggregation

We now aggregate firm-level decisions to obtain implications for sectoral and economy-wide objects. This will highlight how risk compensation distorts aggregate production.

Sectoral and Aggregate Output. Appendix B derives the following expression for expected sectoral GDP in equilibrium:

$$\mathbb{E}[Y_j] = \mathbb{E}[P_j Q_j] = A^* \ \frac{\eta - 1}{\eta} \left(\int (1 - \tau_{ij})^{\eta - 1} A_{ij}^{\eta - 1} \, di \right)^{\frac{1}{\eta - 1}} \qquad \text{where} \tag{15}$$

$$A^* \equiv \left(\int B_j^{\frac{1}{\theta}} A_j^{\frac{\theta - 1}{\theta}} \, dj \right)^{\frac{\theta}{\theta - 1}} , \tag{16}$$

is aggregate productivity. Thus, risk compensation enters as a wedge that reduces effective productivity. Intuitively, risk induces firms to contract production, relative to the level implied by productivity alone.

Since sectors are all symmetric, aggregate output is equal to the expected sectoral GDP derived above, i.e.

$$Q = \mathbb{E}[Y_j] \, . \tag{17}$$

To quantify the output losses from risk considerations, we define the no-risk benchmark Q_j^*, the expected sectoral GDP when $\gamma_i = 0$ (so $\tau_{ij} = 0$):

$$\mathbb{E}[Y_j^*] = A^* \ \frac{\eta - 1}{\eta} \ \left(\int A_{ij}^{\eta - 1} \, di \right)^{\frac{1}{\eta - 1}} . \tag{18}$$

The expected GDP loss due to risk, denoted $\mathcal{L}^{Risk}$, is then given by:

$$\mathcal{L}^{Risk} = \ln \mathbb{E}[Y_j^*] - \ln \mathbb{E}[Y_j] = \frac{1}{\eta - 1} \left[\ln \left(\int A_{ij}^{\eta - 1} \, di \right) - \ln \left(\int (1 - \tau_{ij})^{\eta - 1} A_{ij}^{\eta - 1} \, di \right) \right] \tag{19}$$

Importantly, the aggregate productivity term, A^*—which collects the cross-sector elasticity θ together with the demand and sectoral-productivity shocks $\{B_j, A_j\}$—enters multiplicatively and is common to the baseline and no-risk economies. It therefore cancels in the output loss from risk: the cost of diversification is governed entirely by the within-sector wedge $(1 - \tau_{ij})$ and is independent of θ.

In the next subsection, we outline how we take the model – specifically, equations (15)-(30) – to the data and quantify the effects of risk on markups and output.

2.4 From Model to Data

Sufficient statistics. Firm–sector productivities $\{A_{ij}\}$ and risk compensation, $\{\tau_{ij}\}$ are not directly observable. Using equations (15)-(30), we can recover them from a small set of sufficient statistics: (i) expected revenues by firm-sector $\{R_{ij}\}$, (ii) sectoral demand-shock correlations ρ_{jk}, and (iii) firm-level markups. The model equations directly link these variables to the structural parameters that drive risk-adjusted markups and aggregate distortions.

Measuring risk exposure. Given expected revenue shares and the correlations, risk exposures (both at the firm-sector level and at the firm level) can be directly pinned down from equations (12) and (14).

Identifying the price of risk. Equations (13)-(14) imply that, as a firm expands or contracts its sectoral scope, its risk exposure changes. The price of risk $RiskPr_i$ then governs the change in markups. Therefore, the variation in firm-level markups as risk exposure (or equivalently scope) changes is informative about how much risk matters to the firm. In other words, observed joint movements in firm scope and markups identifies the price of risk.

Recovering productivities. Here, we make use of the model equations which link revenues at the firm-sector level to the corresponding productivities $\{A_{ij}\}$ and markups. Using the risk exposures and price of risk from the steps above, we explicitly account for can markups, allowing us to back out productivities $\{A_{ij}\}$. Specifically, using the firm's optimality conditions and the demand functions, we can derive the following relationship:

$$\ln A_{ij} = \frac{1}{\eta - 1} \ln s_{ij} + \ln\left(\frac{1}{1-\tau_{ij}}\right) + \ln\left(\frac{\eta}{\eta - 1}\right) + \text{Const} \qquad \text{where} \tag{20}$$

$$s_{ij} := \frac{P_{ij}Q_{ij}}{P_j Q_j} \, . \tag{21}$$

denotes firm i's market share in sector j.

To summarize, the approach proceeds in three steps: (i) combine revenue shares with sectoral correlations to measure risk exposure, and (ii) use the relationship between scope and markups to estimate the price of risk and (iii) use revenues to infer market shares and effective productivities. These objects are sufficient to quantify how scope and risk shape markups and matter for

aggregate output. In the following section, we detail our sources and measurement approach.

3 Data and Measurement

3.1 Data sources and definitions

This subsection lays out our data sources and definitions—sector definitions, the firm sample, and markups. The next two subsections then construct the two inputs the model requires: the matrix of sectoral correlations and firm-sector revenue shares.

Sector definitions. We define sectors at the 4-digit North American Industry Classification System (NAICS) industry codes. This level of aggregation balances sufficient granularity with adequate data coverage across our sources.

Firm sample. Our sample consists of publicly traded firms in the US that file annual reports (10-Ks). The sample period runs from 2000 to 2023, comprising approximately 32,687 observations. To capture meaningful changes in firm scope and exposure, we analyze filings at a horizon of $h = 5$ years, pairing 10-K reports five years apart for the same firm. This horizon smooths short-run noise and focuses on medium-run adjustments in business composition.

Markups and risk measures. Firm-level markups are computed using the methodology of De Loecker et al. (2020), implemented on Compustat financial statement data. This approach recovers markups by combining cost shares with estimated output elasticities and is widely used in the empirical industrial organization and macro-productivity literature.[5]

3.2 Measuring Sectoral Correlations

A key input for measurement is the correlation matrix of $\{X_j\}$, the composite sectoral shock. To estimate this, we aggregate the optimal labor choice from (8) to the sectoral level (with time subscripts):

$$L_{ijt} = \left(\frac{1}{W_t}\frac{\eta-1}{\eta}\right)^{\eta} \mathbb{E}_{it}[X_{jt}](1-\tau_{ijt})^{\eta} A_{ijt}^{\eta-1} ,$$
$$\ln L_{jt} = \ln\left(\frac{1}{W_t}\frac{\eta-1}{\eta}\right)^{\eta} + \ln \int \mathbb{E}_{it}[X_{jt}](1-\tau_{ijt})^{\eta} A_{ijt}^{\eta-1} di .$$

We assume (i): $\ln X_{jt}$ is conditionally normal and follows a first-order autoregressive process,

[5] We use an open-source Python implementation published by Basile Grassi to ensure consistency and replicability.

i.e. $x_{jt} \equiv \ln X_{jt} = \varrho_x x_{j,t-1} + e_{jt}$ with $\mathbf{e}_t \sim N(0, \Sigma)$ and (ii): all firms make decisions based on the commonly observed history of past realizations, i.e. the information set is $\mathcal{I}_{it} = \{X_{jt-s}\}_{s=1}^{\infty}\ \forall i$. Then,

$$\begin{aligned} \ln E_{it}[X_{jt}] &= E_{it}[\ln X_{jt}] + \text{Constant} = \varrho_x x_{j,t-1} + \text{Constant} . \\ \Rightarrow \ln L_{jt} &= \ln\left(\frac{1}{W_t}\frac{\eta-1}{\eta}\right)^{\eta} + \varrho_x x_{j,t-1} + \ln \int (1-\tau_{ijt})^{\eta} A_{ijt}^{\eta-1} di + \text{Constant} . \end{aligned}$$

Finally, we also impose a stationarity assumption: the penultimate term (or equivalently, the joint distribution of wedges and productivity) is assumed to be constant over time (or changing in a symmetric way across sectors). Under these assumptions, the co-movement of sectoral employment (net of a time fixed effect to remove common components) recovers the correlation matrix of sectoral risk.

In our empirical implementation, we use data on annual employment for 288 4-digit NAICS sectors from the Business Dynamics Statistics (BDS), published by the U.S. Census Bureau for the period 1978-2022.

3.3 Measuring Revenue Shares

To compute a firm 's risk exposure, we need to know not only the sectors in which it operates (what previous measures captured), but also the revenue shares of each of these sectors. Traditional sources for firm-level data do not directly yield revenue shares by sector. The 10-k also contain details on business segments, but that information is too coarse for our purposes. The definition of segments does not necessarily correspond to sectors. Many firms (for example, Apple and Walmart) report detailed information only by geographical segment. Moreover, a significant firms (as many as 80% in the more recent years) only report information for 1 business segment. Another potential source is administrative data (e.g. from the US Census). However, information in the US Census data is limited in at least 2 respects. First, it is restricted to domestic operations. Second, it does not allocate the revenue of multi-sector establishments (e.g. headquarters).

A third approach uses textual analysis of regulatory filings. Prior work, notably Hoberg and Phillips (2025), measures a firm's industry exposure as the fraction of words in its 10-K that match industry-specific dictionaries. While innovative—and useful for capturing features of a firm's operations—this approach is designed to identify the extensive margin of operation, whether a firm is active in an industry, rather than the revenue *shares* we require. Mapping such word-

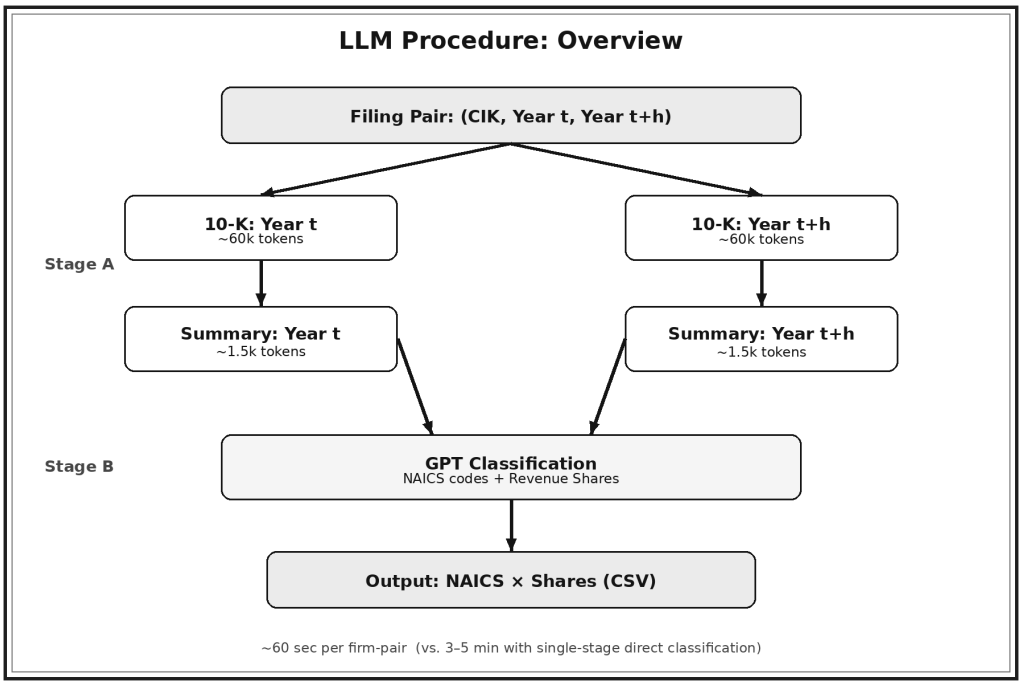


Figure 1: Two-stage LLM pipeline to construct firm-sector revenue shares from 10-K filings

based exposure into firm-level revenue shares requires strong assumptions, and small changes in language can generate noisy inferences about economic activity.

Therefore, we devise a new approach to identify sectors and measure of firm-sector revenue shares (π_{ijt}). We use a large language model (LLM) to parse the business, risk factors, and management's discussion and analysis sections of firm 10-K filings. We collect all the 10-K filings from the U.S. Securities and Exchange Commission using the SEC-API between 2000 and 2025. The approach relies on narrative disclosures in firms' annual 10-K filings. Specifically, we use ChatGPT's GPT-5.2 model to translate qualitative business descriptions into quantitative sectoral revenue shares, which can be mapped directly to our model-based measures of firm scope and risk exposure. Figure 1 summarizes the two-stage pipeline.

To aid replicability and minimize simulation noise, we use a two-stage procedure. The first stage summarizes the relevant business-line descriptions from Sections 1, 1A, 1B, 7 and 8. The model reads selected sections of the 10-K and produces a summary that isolates information relevant for industry classification. The prompt for stage 1 input reads as follows:

Stage A Prompt

Analyze this 10-K filing and extract ONLY revenue-relevant information for NAICS classification.

OUTPUT FORMAT - Provide a structured summary with these sections:

1. BUSINESS SEGMENTS: List each business segment/division with:

- Segment name; Products/services offered
- Estimated revenue share (% of total) if mentioned

2. REVENUE SOURCES: List primary revenue streams:

- Product sales vs services; Geographic breakdown if available
- Customer types (B2B, B2C, government, etc.)

3. INDUSTRY INDICATORS: Key phrases indicating industry:

- Manufacturing activities; Retail/wholesale operations
- Technology/software services; Financial services; Healthcare activities

4. CHANGES NOTED: Any mentions of:

- New business lines; Discontinued operations; Acquisitions or divestitures

Keep the summary concise (under 2000 words). Focus on facts that help determine 4-digit NAICS codes.

The sections used are those most likely to contain economically meaningful descriptions of the firm's operating activities: Section 1 (Business) describes the firm's main products and markets, Section 1A (Risk Factors) discusses key risks facing the firm; Section 1B (Unresolved Staff Comments) contains regulatory commentary when relevant; and Section 7 (Management's Discussion and Analysis) provides management's perspective on operations and performance. Together, these sections provide rich descriptions of the firm's economic activities across sectors. The output is a concise structured summary, that preserves the parts of the filing most informative for mapping the firm into 4-digit NAICS industries. Because we can store the descriptions, this stage allows others to use these summaries to replicate the analysis, despite changing LLM models and outputs. The first stage also reduces noise by focusing the model's attention on text that contains specific information about revenue-generating activities and business segments, avoiding the inclusion of unrelated industrial codes.

In the second stage, we pass these summaries to GPT-5.2, to obtain approximate revenue

shares. We pass two filing summaries for the same firm at a five-year horizon ($h = 5$) and ask the model to produce a structured CSV output containing all active 4-digit NAICS codes and the corresponding revenue shares in the base and final years. The prompt requires the model to list all NAICS codes appearing in either year, assign a revenue share to each code in both periods, using zeros when a code is absent in one year. Stage B prompt reads as follows:

Stage B Prompt

Based on these revenue summaries from two years, identify the 4-digit NAICS codes the company operates in. REQUIREMENTS:

1. List all NAICS codes from both years combined
2. For each code, provide revenue share (%) for Year 1 (Base) and Year 2 (Final)
3. If a code wasn't present in a year, use 0 for that year's share
4. CRITICAL: Base Share column MUST sum to exactly 100
5. CRITICAL: Final Share column MUST sum to exactly 100

Output format (CSV):
NAICS Code, Base Share, Final Share
3361, 50, 10
5413, 45, 40
2389, 0, 30
3399, 5, 20
TOTAL, 100, 100
Include the TOTAL row to verify your sums equal 100.

By passing both years jointly, we force the model to reason about the evolution of the firm's business mix across filings. This paired-year setup helps the algorithm distinguish true business-line expansion, contraction, entry, and exit from random variation in wording across annual reports. We also ask the model to ensure that the shares in each year sum to exactly 100. [6] The final

[6]The output is accepted only if the revenue shares for each year sum to 100 up to a small tolerance of 5%. Failed instances are reprocessed with stronger instructions.

output has the following form: NAICS code, base share, final share. The procedure yields a mapping from business line descriptions in firm 10-Ks directly to 4-digit NAICS sectors. Combined, the resulting revenue shares can be integrated to reveal sectoral distances.

Table 1 reports descriptive statistics for the main variables in our analysis. Panel B summarizes the estimated revenue shares of multi-sector firms (those operating in more than one sector): the average multi-sector firm operates in 3.86 sectors but maintains a relatively concentrated portfolio, with its 'top' sector accounting for 76% of revenues on average and the top three for over 95%. Panel A reports firm size and markups: log sales are widely dispersed (interquartile range 3.9–7.1), and the median markup is 1.39 (interquartile range 1.13–1.97). Panel C describes sectoral structure: a typical sector contains about 23 firms and is fairly concentrated, with an average Herfindahl index (HHI) of 0.42.

Table 1: Descriptive Statistics

Variable	Mean	Std. Dev.	P25	P50	P75
Panel A. Firm scope, size, and markups					
Scope	0.18	0.23	0.01	0.08	0.28
Log sales	5.33	2.62	3.87	5.48	7.09
Markup	1.86	1.44	1.13	1.39	1.97
Panel B. Revenue composition (multi-sector firms)					
Top-1 sector share	0.76	0.20	0.60	0.81	0.93
Top-2 sector share	0.16	0.13	0.05	0.12	0.25
Top-3 sector share	0.05	0.06	0.00	0.02	0.08
No. of sectors	3.86	2.17	2.00	3.00	5.00
Panel C. Sectoral structure					
No. of firms	23.1	39.3	5.0	10.0	22.0
Top-firm share	0.55	0.23	0.37	0.52	0.72
HHI	0.42	0.24	0.23	0.37	0.56

Notes: Panels A and B report firm-year moments and Panel C reports sector-year moments. The firm-level sample contains 33,655 firm-years (markups are available for 22,737); Panel B is computed on the multi-sector subsample (firms operating in more than one sector). The sectoral sample contains 4,958 sector-years; a sector-year with fewer than three firms is merged with its nearest sector in distance, with the combined-sector averages assigned to both.

Validation

We conduct two experiments aimed to instill confidence in the LLM-generated revenue shares. First, for selected firms, we manually checked the NAICS codes and shares assigned by the LLM

against the textual evidence in the underlying 10-Ks, including segment disclosures where available. These checks are not intended to claim that the LLM recovers true revenue shares without error. Rather, they verify that the estimated shares are grounded in recognizable business descriptions and that the inferred changes over time align with well-known shifts in firms' product portfolios. Table 2 shows Amazon's business line over the last two decades as captured by the LLM. In 2000, its 10-K describes a firm that is overwhelmingly an online retailer, and the model correspondingly assigns almost all revenue to online retail. By 2020, the filing emphasizes not only retail, but also third-party seller services, AWS cloud computing, physical grocery operations, and advertising. The model translates these changes in narrative prominence into a more diversified NAICS revenue vector, with retail remaining largest but substantial shares emerging in cloud, marketplace services, grocery, and advertising.

These manual checks also show what the LLM adds relative to the standardized segment data, which we argued above is too coarse for our purposes: about 55% of the firm-years in our sample with a Compustat business-segment record report a single segment covering the entire firm. Apple is a stark example. In fiscal 2024, Compustat records a single business segment spanning all $391 billion of Apple's revenue, with disaggregation only by geography. The LLM instead recovers the product-line split disclosed in the narrative of the filing almost exactly: its five 4-digit NAICS shares are each within one percentage point of the reported revenue shares of iPhone, Services, Wearables, Mac, and iPad. The pattern is general: in roughly a third of the single-segment firm-years, the LLM identifies a second sector accounting for at least 10% of revenue, information necessarily extracted from the text since the segment data for these firms is a single line covering the whole firm. Appendix D provides details.

Our second validation exercise asks how much of a firm's sectoral composition is recoverable from the 10-K narrative alone, when the filing's quantitative segment table is not available to the model. We run two versions of Stage B that differ only in their input. *Version 1* (text-only) receives the Stage A narrative summary by itself. *Version 2* (segment-augmented) receives the same summary together with the firm's reported business-segment revenue data. The two versions are otherwise identical—same model, same prompt, same firm-pairs. The experiment is informative only where the narrative summary itself lacks quantitative segment figures, since elsewhere both versions see essentially the same information. We therefore restrict attention to the 2,789 firm-pairs (the *test sample*, about 11% of the full panel of firm-pairs) whose Stage A summaries contain essentially no quantitative segment information and for which segment data was available to supply in

Table 2: Amazon NAICS Assignment and Revenue Share Validation

Panel A. Amazon 2000		
NAICS sector	**Share**	**Validating Evidence (Human Audit)**
Online retail (NAICS 4541)	98%	**Sem:** "leading online retailer"; "online retail stores ... books, music, DVD/video". **Int:** Retail is presented as the core business, with the longest and richest product discussion, implying a dominant share.
Online marketplaces (NAICS 5191)	2%	**Sem:** "Web-based marketplaces ... transactions"; "Auctions, zShops". **Int:** Positioned as a new add-on with limited detail, implying a small share.
Panel B. Amazon 2020		
NAICS sector	**Share**	**Validating Evidence (Human Audit)**
Online retail (NAICS 4541)	57%	**Sem:** "online and physical stores"; "hundreds of millions of unique products". **Int:** Retail and fulfillment are framed as the core consumer business, implying the largest share.
Marketplace / nonstore retail (NAICS 4543)	20%	**Sem:** "sellers... sell their products in our stores"; "not the seller of record... earn fees". **Int:** Seller programs are sizeable, but clearly secondary to first-party retail.
Data processing / cloud (NAICS 5182)	12%	**Sem:** "AWS segment... global compute, storage, database". **Int:** Segment with repeated emphasis, implying a large services share.
Grocery / physical stores (NAICS 4451)	6%	**Sem:** "Whole Foods Market... included... from acquisition"; "physical stores". **Int:** Physical grocery is present with moderate discussion, implying a modest share.
Advertising (NAICS 5418)	5%	**Sem:** "Other" revenue "primarily includes sales of advertising services". **Int:** Advertising is the principal component of the fast-growing "Other" line, implying a small but rising share.

Panel A note: The 2000 classification is primarily retail and is supported by actual segment revenue disclosure in the 10-K: U.S. retail contributes $1.3 billion out of $1.6 billion in overall revenue.

Panel B note: The 2020 classification is supported by disclosed segment information: AWS accounts for about 12% of net sales, Whole Foods enters the filing after the 2017 acquisition, and advertising is reported within the fast-growing "Other" revenue line.

Evidence columns: "Sem" reports semantic evidence quoted or paraphrased from the filing; "Int" reports our interpretation of how prominently the activity features in the narrative.

version 2. On the test sample, version 1 must infer the firm's sectoral composition from narrative alone, while version 2 has the segment figures as an anchor, so comparing the two isolates the information content of the text.

The two versions agree closely. Version 1 identifies the same top sector as version 2 in 76% of the test sample at the 4-digit NAICS level and in 87% at the 2-digit sector level. Comparing the full revenue-share vectors, the mean and median cosine similarities between the two versions are 0.79 and 0.99, respectively.[7] The disagreement is concentrated in which sectors are listed rather than in how revenue is allocated across them: conditional on the two versions selecting the same set of sectors, the mean cosine similarity rises to 0.98. The two approaches therefore produce very similar measures of scope.

We further examine the tail of cases where the two versions disagree most—the 19% of the test sample with cosine similarity below 0.5—to ask whether these reflect genuine misclassification or merely benign differences in coding. Classifying each disagreement by the relationship between the two versions' top sectors, we find that most are not errors. Well-documented NAICS taxonomy overlaps, for which multiple four-digit codes are defensible for the same activity (for example, pharmaceutical R&D vs. pharmaceutical manufacturing, or packaged software vs. data-processing and hosting), account for 5.3% of the test sample, and granularity refinements that agree on the three-digit industry but differ in the four-digit sub-code account for a further 4.4%. The disagreements that differ at the broad two-digit sector level—6.4% of the test sample—are the only candidates for true error. For these we use Compustat's independent firm-level NAICS code as a tie-breaker: it sides with the segment-augmented version in 1.8% of the test sample, with the text-only version in 0.5%, selects a third (still defensible) code in 2.2%, and is unavailable in 1.2%. Treating the cases where Compustat contradicts the text-only version as confirmed errors implies a text-only error rate of 1.8% of the test sample. These confirmed errors share a common pattern: the firm is identified by the activity its 10-K emphasizes rather than by its finished output. For instance, when the narrative dwells on selling or distributing a product, the text-only model classifies the firm as a wholesaler or retailer even though it manufactures that product—a distinction

[7] For each firm-pair and each version we construct a vector of final-year revenue shares indexed by 4-digit NAICS sector, over the union of sectors that appear in either version (a sector absent from one version enters with a zero share). The cosine similarity is the dot product of the two share vectors divided by the product of their Euclidean norms, $\cos(s^1, s^2) = \frac{s^1 \cdot s^2}{\|s^1\| \, \|s^2\|}$. It equals one when the two versions assign identical shares to identical sectors and zero when their sector sets are disjoint, and it reflects agreement both in which sectors are present and in how revenue is allocated across them. The same-set statistic restricts to firm-pairs in which both versions list exactly the same set of sectors, so that the comparison reflects only differences in the shares.

that the segment data, and Compustat's independent firm-level code, both pick up. Overall, the disagreements that survive scrutiny are rare and systematic rather than pervasive noise.

Together, these exercises provide a measure of validation for the LLM's ability to extract sectoral composition from the text of the filings.

4 Recovering Sufficient Statistics

Section 3 provides the two inputs the theory requires—firm-sector revenue shares and the matrix of sectoral correlations. We now combine them to construct and estimate the three objects that serve as sufficient statistics for quantifying the output loss from diversification motives, in the order the model builds them: firm *scope*, the *price of risk*, and firm-sector *productivities*. Scope depends only on the inputs already in hand; the price of risk is then estimated from scope and measured markups; and productivities are recovered using the estimated price of risk. Each object is motivated by the theory of Section 2, and we validate each against the data as we proceed.

4.1 Scope

Our measure of scope is grounded in the model's notion of risk exposure. Recall that firm i's risk from its operations in sector j depends on how revenues in that sector comove with the firm's total revenues. This logic naturally delivers a risk-based measure of scope: firms are broader in scope when their revenue composition provides diversification benefits and narrower in scope when their activities are concentrated in correlated sectors.

Empirically, we combine two ingredients: (i) firm-sector revenue shares we obtain from our LLM procedure and (ii) sectoral correlations. The model defines scope in (14) as a function of *expected* revenue shares $\{R_{ijt}\}$. The measurement strategy in the previous subsection yield *realized* revenues. Let $\frac{Rev_{ikt}}{Rev_{it}}$ denote the realized share of firm i's revenues in sector k in period t. Plugging into (14) yields the measured scope of firm i at time t is:

$$Scope_{it} = \sum_{j,k} \frac{Rev_{ijt}}{Rev_{it}} \frac{Rev_{ikt}}{Rev_{it}} (1 - \rho_{jk}). \tag{22}$$

Recall from equation (14) that $Scope_{it} = 1 - RiskExp_{it}$. While risk exposure is the object that enters the markup equation (13), scope is arguably more useful for building intuition and for connecting to the broader literature on diversification.

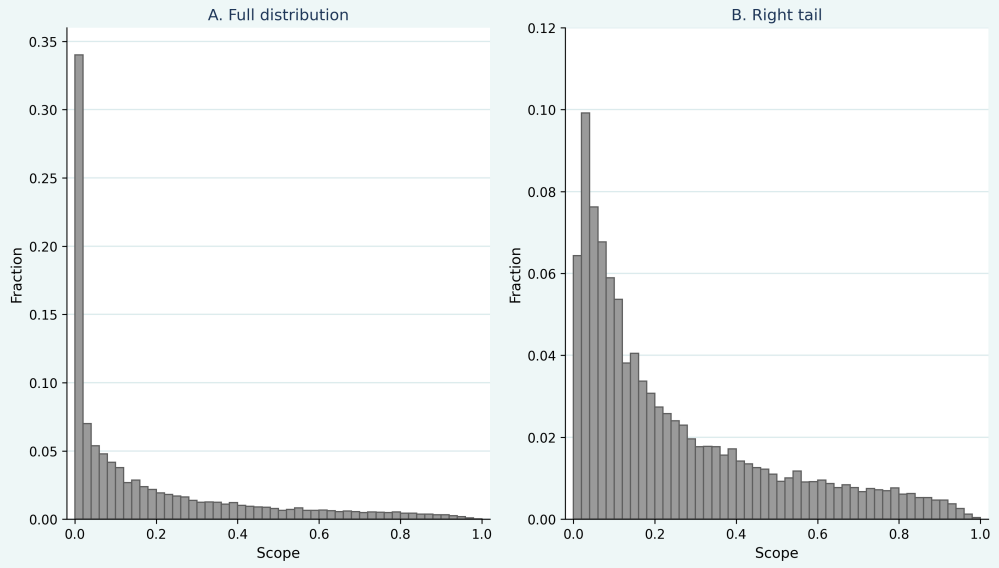


Figure 2: Histogram of Firm Scope Measure

Notes: Panel A shows the full distribution of Scope. Panel B excludes observations with Scope $\leq$ 0.01 *in order to zoom in on the right tail of the distribution.*

Figure 2 depicts a histogram of our scope measure. It shows a significant mass at zero (single-sector firms) and a long right tail of highly diversified conglomerates. Across firms, scope averages 0.18, with an interquartile range of 0.01 to 0.28 (Panel A of Table 1).

One might reasonably doubt that applying a language model to firm 10-Ks would deliver sensible revenue shares and scope measures. As an external check on the measure, we ask whether scope predicts firm exit. In Table 3, we run a probit regression where the dependent variable equals 1 if the firm exits within the next 5 years (columns 1–2) or 10 years (columns 3–4) and 0 otherwise. Exit is defined as a firm permanently leaving the Compustat sample. A firm that last appears in Compustat at t is said to exit at t. The independent variables are standardized values of $\ln(1 + Scope)$ and controls. The results show that the scope has a statistically and economically meaningful association with exit. A 1-standard deviation change in $\ln(1 + Scope)$ is associated with a 5% reduction in the probability of exit over the next 5-10 years. This pattern is consistent with the hypothesis at the core of this paper: expanding scope reduces risk.

4.2 The Price of Risk

Next, we estimate the relationship between firm-level markups and scope. The theory implies that firms with narrower effective scope face more undiversified revenue risk and therefore charge

Table 3: Scope and Exit.

	P(Exit in 5yrs)	P(Exit in 5yrs)	P(Exit in 10yrs)	P(Exit in 10yrs)
log(1+scope)	-0.0508***	-0.0505***	-0.0589***	-0.0625***
	(-6.36)	(-5.84)	(-6.90)	(-6.72)
log inf adj. Rev	-0.2476***	-0.2559***	-0.2904***	-0.3168***
	(-30.07)	(-27.55)	(-32.03)	(-28.74)
Observations	26758	19976	23826	17838
Year FE	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes

t statistics in parentheses

Notes: The dependent variable is an indicator equal to one if the firm exits within the next 5 years (columns 1–2) or 10 years (columns 3–4). All independent variables are standardized as $(x - \bar{x})/\sigma_x$. Additional controls (included where indicated) are R&D intensity, operating profits, and net leverage. All specifications include year fixed effects. Standard errors are robust.

higher expected markups as compensation for that risk. This relationship will also help us pin down a key parameter: the price of risk.

The model delivers a structural relationship between two key objects: the expected firm-level markup and firm scope. Let μ_{it}^* denote the model-implied expected markup of firm i in year t, and let Scope_{it}^* denote the model scope constructed from expected firm-sector revenue shares. Equations (13)-(14) imply that these objects satisfy

$$\mu_{it}^* = \frac{\eta}{\eta - 1} \frac{1}{1 - \varphi(1 - \text{Scope}_{it}^*)}. \tag{23}$$

where $\varphi \equiv 2\hat{\gamma}\sigma^2$ is a composite parameter that bundles the price of risk with the variance of sectoral revenue shifters (in logs) and is the single object that governs risk-adjusted markups. This uses the assumption that the decision-maker's absolute risk aversion declines with firm size, $\gamma_{it} \propto 1/R_{it}$, so that the product $\gamma_{it}R_{it} = \gamma$—and hence the price of risk $2\gamma_{it}R_{it}$—is constant across firms and time.

We operationalize equation (23) using data counterparts of the model objects. Let μ_{it} denote the measured firm-level markup and let Scope_{it} denote the measured firm-level scope variable constructed using the observed revenue shares from 10-Ks. The markup μ_{it} is subject to yearly 1/99 winsorization. We compute the measured markup following the production-function approach of De Loecker et al. (2020).

Identification comes from within-firm movements in measured scope and measured markups. Let

$$\widetilde{\text{Scope}}_{it} = \text{Scope}_{it} - \overline{\text{Scope}}_i, \qquad \widetilde{u}_{it}(\varphi) = u_{it}(\varphi) - \overline{u}_i(\varphi).$$

where

$$u_{it}(\varphi) = \mu_{it} - \frac{\eta}{\eta - 1} \frac{1}{1 - \varphi(1 - \text{Scope}_{it})}$$

The moment condition is

$$\mathbb{E}\left[\widetilde{\text{Scope}}_{it} \widetilde{u}_{it}\right] = 0. \tag{24}$$

We follow the firm dynamics literature and set $\eta = 5$ – see, for example, Atkeson and Burstein (2010).

The first column of Table 5 show the estimated φ is both statistically and economically significant. To see the latter, note that, at the mean measured scope in our sample, $\overline{\text{Scope}} = 0.18$, the estimated $\varphi = 0.150$ implies a markup wedge of

$$\frac{1}{1 - 0.150\,(1 - 0.18)} = 1.14.$$

In other words, the risk wedge scales the standard CES markup $\frac{\eta}{\eta - 1}$ by a factor of 1.14. Thus, this evidence on markups supports our central premise: un-diversified risk has a substantial effect on production and markups.

The remaining two columns in Table 5 address two potential concerns with our estimation strategy. The first has to do with the distinction between expected and realized revenue shares. Equation (23) links *expected* markups to scope, both constructed from expected revenue, whereas the observed markups and scope reflect are based on actual, or *realized* revenue. As a result, a contemporaneous sectoral shock can create a mechanical correlation between scope and markups. The sign of this correlation (and therefore, the bias in our estimate) is ambiguous – it depends on the risk exposure of the sector subject to the shock, relative to the average risk exposure of the firm has a different risk exposure than the firm's overall portfolio.

To address this concern, we re-estimate φ by instrumenting current scope with lagged measured scope (since we sample firms' 10-K at 5-year intervals, this corresponds to scope from 5 years ago). The moment condition in this IV-GMM specification is

$$\mathbb{E}\left[\widetilde{\text{Scope}}_{i,t-5} \widetilde{u}_{it}\right] \ . = 0 \tag{25}$$

The identifying assumption is that lagged measured scope predicts current scope but is orthogonal to current sectoral shocks. Table 4 shows a strong first stage: lagged scope has a sizable relationship with the current measured scope.

Table 4: First Stage for Lagged-Scope IV

Statistic	Value
Lagged scope coefficient	0.220
	(0.016)
Firm-clustered Wald F-statistic	181.071
Within R^2	0.057
Observations	14,786

Notes: The dependent variable is $\widetilde{\text{Scope}}_{it}$. The regressor is $\widetilde{\text{Scope}}_{i,t-5}$. Both variables are demeaned within firm in the lagged-scope IV sample. Firm-clustered Standard error reported in parentheses.

The second column of Table 5 reports the results from this specification. It shows that the estimated φ is very close to that obtained under the baseline GMM specification. [8] This suggests that the confounding effect of contemporaneous shocks is small, perhaps due to the ambiguous sign. We use $\hat{\varphi} = 0.150$ from the within-firm baseline as our preferred estimate in the rest of the paper.

Table 5: Scope and Markups: Price-of-Risk Estimates

	Baseline	IV	Size-Dep
$\hat{\varphi}$	0.150	0.157	0.147
	(0.017)	(0.074)	(0.018)
Market share coeff			0.0007
			(0.004)
Observations	22,207	14,786	22,207

Notes: The table reports GMM estimates of $\varphi \equiv 2\gamma\sigma^2$, the parameter governing the markup–scope relationship in equation (23) (measured markup μ_{it} and measured scope Scope_{it} replace their model counterparts, with $\eta = 5$ and markups yearly 1/99 winsorized). The "Baseline" column is the within-firm GMM estimator on the full sample (22,207 firm-years, 7,421 firms), and "IV" is the lagged-scope IV GMM estimator on the lagged-IV sample (14,786 firm-years, 6,577 firms; median gap of five years between current and lagged scope). The *Size-dependent* column also lets the sector-level demand elasticity load on the firm's log within-sector market share. Bootstrapped standard errors in parentheses. Appendix **??** develops the size-dependent specification, reports bootstrap standard errors, and gives an IV version.

The last column addresses another potential issue: markups that vary systematically with size- (or more precisely, market shares). To see why this is a concern. Consider a firm that experiences an increase in idiosyncratic productivity in its home sector. This raises the revenue weight of the home sector and would, *ceteris paribus*, reduce scope. But it also raises the firm's market share in that sector. In many models of variable markups, this would increase markups: in other words, markups rise and scope falls even in the absence of risk consideration.

[8] Since instrumenting requires the lagged scope, it shrinks the sample relative to the baseline. But this change in sample does not play a significant role in the results: running the baseline specification on the restricted sample produces an estimate of 0.148, almost unchanged from the baseline value of 0.150.

To address this, we re-estimate φ explicitly allowing for the firm's baseline markup in a sector to vary with its market share. We relegate the details to Appendix **??** and report the estimates – for both φ and the loading on market shares – in the last column of Table 5. They show the price of risk is essentially unchanged, $\hat{\varphi} = 0.147$, with $\hat{\xi}$ indistinguishable from zero.

4.3 Productivity

To quantify the gains to specialization, we measure the change in a firm's total factor productivity as it moves away from its core competency, defined as the 4-digit NAICS industry in which the firm achieves its maximum measured productivity (A_{ijt}). This home sector, denoted j_i^*, serves as the benchmark: the distance of a given sector k from firm i's core competency is defined as $1 - \rho_{kj_i^*}$.

The estimation proceeds in two steps. First, we non-parametrically estimate a hazard function to determine the probability that a firm operates in a given sector as a function of the distance, $1 - \rho_{kj_i^*}$. Second, we recover firm-sector productivity (A_{ijt}) from sectoral market shares (or more precisely, the relative size of a firm in a sector), price of risk (estimated above) and $RiskExp_{ij}$ (measured as revenue weighted correlation of sector j to all sectors firm i operates in), leveraging the insight in Equation (20). The sectoral market share of a firm is computed as the firm's revenue in a sector relative to the average revenue of all firms operating in that sector.[9]

Figure 3 displays the estimated hazard: it shows the average probability of a firm operating in a sector for quantile bins of the distance $1 - \rho_{kj_i^*}$, along with a fitted line:

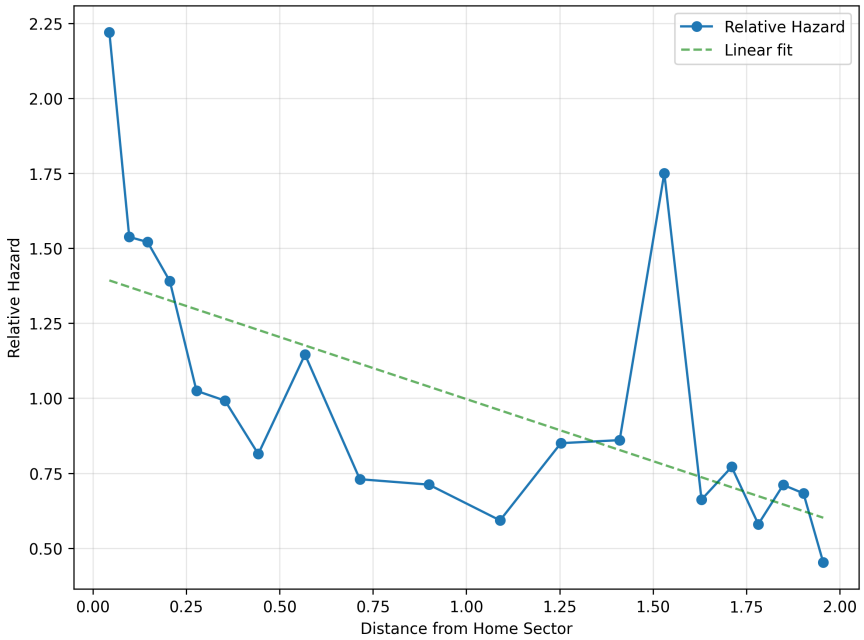
$$h(\rho_{kj^*}) = 1.41 - 0.42(1 - \rho_{kj^*}) \tag{26}$$

where $h(\rho_{jk})$ denotes the likelihood of operating in sector $k \neq j^*$. The estimates imply that the probability declines sharply as the correlation with the home sector falls: a firm is about 2.2 times more likely to operate in a sector whose correlation with its home sector is 0.9 than in a sector whose correlation is -0.9. This extensive margin pattern is inconsistent with theories of firm scope where new products arrive randomly.

Figure 4 depicts the gains to specialization: it plots the average productivity of non-home sector (relative to home sector TFP) against distance $1 - \rho_{kj_i^*}$ (in quantile bins). The estimted line

[9]The model assumes symmetry across sectors. In our empirical implementation, we remove a sector-by-year fixed effect from the estimated A_{ijt} to control for fixed heterogeneity across sectors.

Figure 3: Firms Less Likely to be in Sectors Far from Home



Notes: The figure plots the fraction of firm-years operating in a sector at a given distance from the home sector, using quantile bins of sectoral distance. Relative hazard rescales this fraction by its average across bins.

of best fit implies that the TFP of firm i in a sector $k \neq j_i^*$

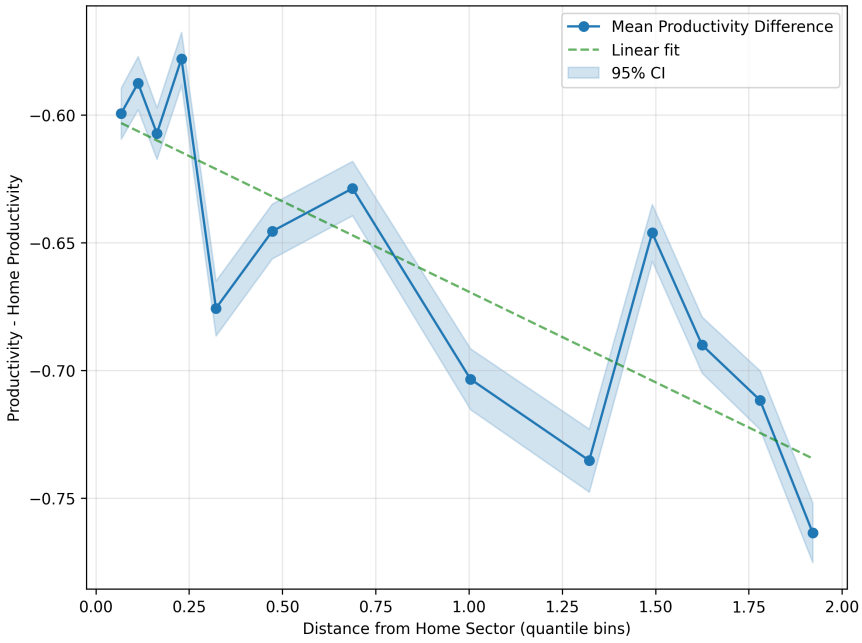
$$\ln A_{ik} = -0.6 - 0.071\,(1 - \rho_{kj_i^*}) \tag{27}$$

Productivity declines as a firm moves away from its core competency: the penalty for a sector close to the home sector is about 60 log points, which increases to 75 log points for sectors that are furthest away. This TFP decay implies that diversifying into distant sectors may provide risk-reduction benefits, but at an efficiency cost.

5 Quantification: The Cost of Uncertainty

Does the diversification vs. specialization trade-off matter for the macroeconomy? We use the estimates so far and the model structure to quantify the output cost of firms' diversification. This is not a welfare cost: Firms optimally balance foregone output and the benefits of risk reduction. To quantify aggregate output losses from risk, we conduct a simple a Lucas (1987) style exercise: we compare the GDP of our baseline economy to a counterfactual economy where the price of risk

Figure 4: Gains to Specialization: Distance from Home Sector Reduces Productivity



Notes: The figure plots the log difference in productivity relative to the firm-year home sector against distance from the home sector, using quantile bins of sectoral distance. Productivity is net of sector-year fixed effects. The home sector is defined as the sector with the highest implied productivity within the firm-year.

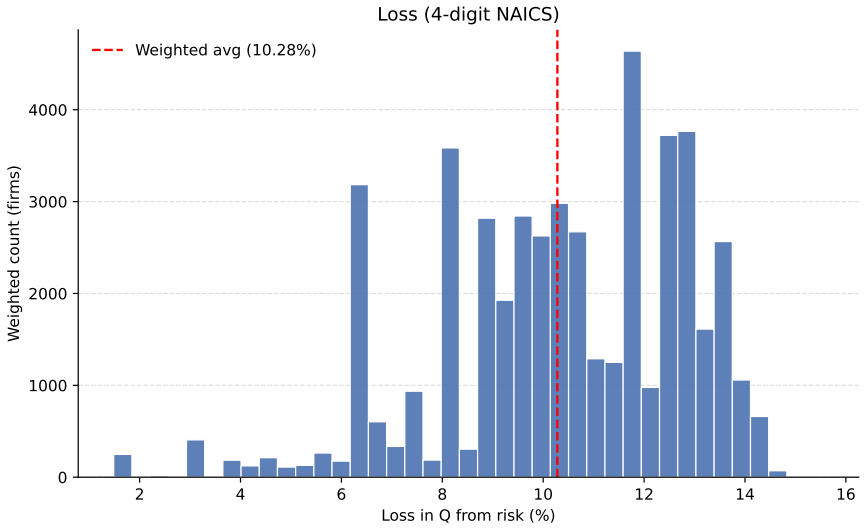
parameter is set to zero ($\gamma = 0$).

We perform this exercise separately for each 4-digit sector in our sample. For each sector j, we use the estimates of productivity and risk wedges from the previous section to construct the loss due to risk from (19):

$$\mathcal{L}_j^{Risk} = \ln \mathbb{E}[Y_j^*] - \ln \mathbb{E}[Y_j] = \frac{1}{\eta - 1} \left[\ln \left(\int A_{ij}^{\eta - 1} \, di \right) - \ln \left(\int (1 - \tau_{ij})^{\eta - 1} A_{ij}^{\eta - 1} \, di \right) \right] . \tag{28}$$

The estimated losses are depicted in Figure 5, which plots a weighted histogram of $\mathcal{L}_j^{Risk}$ (the weights being the number of firms in the sector). It shows large losses across the board, ranging from 0.2% at the low end to 14.6%. The weighted average (weighted by the number of firms in the sector) is 10.28%.

Figure 5: Output Loss from Risk



Notes: The figure plots a histogram (weighted by the number of firms in each sector) of the sectoral output loss due to risk, as defined in (19).

To gain more intuition, we define another useful benchmark: GDP in a counterfactual economy where the distribution of wedges $1-\tau_{ij}$ is the same, but they are independent of productivity, A_{ij}:

$$\mathbb{E}[Y_j^{Ind}] = A^* \ \frac{\eta-1}{\eta}\left(\int (1-\tau_{ij})^{\eta-1} di \ \int A_{ij}^{\eta-1}\, di\right)^{\frac{1}{\eta-1}}. \tag{29}$$

We can then decompose the total loss into two components: one stemming from the average wedge and the other attributable to the covariance between the wedge and productivity:

$$\mathcal{L}_j^{Risk} = \underbrace{\ln \mathbb{E}[Y_j^*] - \ln \mathbb{E}[Y_j^{Ind}]}_{\text{Average Wedge Loss} \equiv \mathcal{L}^{Avg}} + \underbrace{\ln \mathbb{E}[Y_j^{Ind}] - \ln \mathbb{E}[Y_j]}_{\text{Covariance Loss} \equiv \mathcal{L}^{Cov}} \tag{30}$$

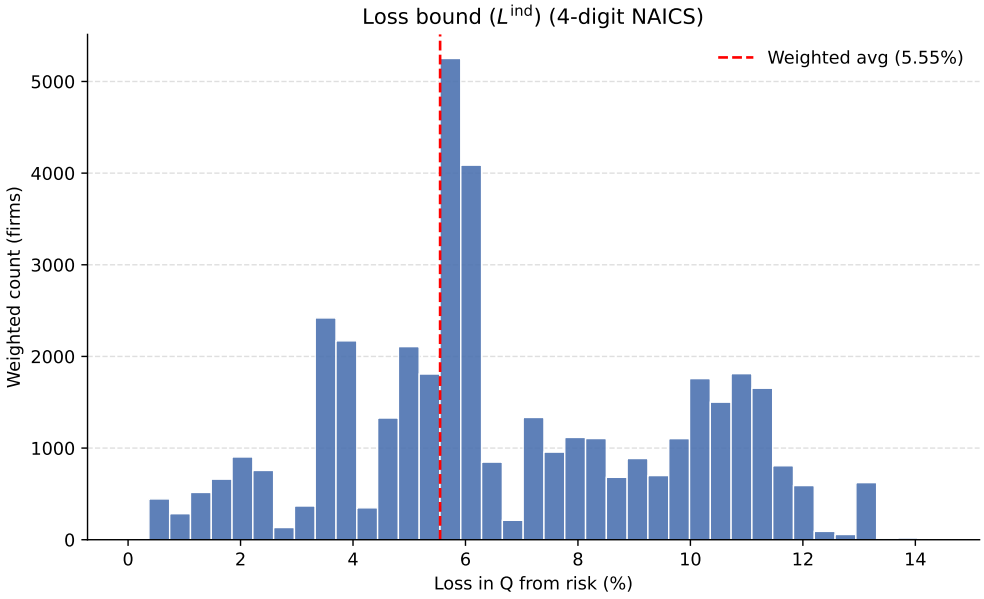
The first term, $\mathcal{L}^{Avg}$, is equal to $\frac{-1}{\eta-1}\ln\int(1-\tau_{ij})^{\eta-1}di$: the reduction in output due to the average level of risk compensation among firms operating in the sector. The second term arises when the risk wedge varies systematically with productivity. When more productive firms require higher risk compensation, this covariance is positive and further depresses output. In contrast, if highly productive firms are better diversified and face lower τ_{ij}, the covariance term can mitigate aggregate losses. Thus, output depends not only on the distribution of productivities but also on

the distribution of risk compensation and its correlation with productivity.[10]

Figure 6 depicts a histogram of the loss attributable to the average wedge, $\mathcal{L}^{Avg}$. The weighted average across sectors amounts to 5.55%, which is a little over half of the total loss from risk depicted in Figure 5. In other words, the covariance of firm-level productivities with the losses from risk significanly amplifies the aggregate consequences from the rsk wedge. Specifically, the fact that the more productive firms in a sector demand higher compensation makes risk more costly.

This pattern is driven partly by the substantial gains from specialization discussed above. *Ceteris paribus*, the higher a firm's productivity in a sector, the greater its risk exposure and therefore, higher is the risk compensation required. This is stronger when productivity decays sharply with the potential for diversification. Intuitively, a higher cost of diversification makes the high productivity firms in a sector more exposed to risk, driving up the correlation between risk and productivity (and through that, the aggregate output losses from risk). In the following subsection, we use a simple example to formally show this connection.

Figure 6: Sectoral Output Loss from Average Wedge, $\mathcal{L}^{Avg}$



[10] Another way to see this decomposition is to approximate (17) around the no-risk benchmark:

$$\ln \mathbb{E}[Q_j] \approx \ln \mathbb{E}[Q_j^*] - \left[\overline{\tau_j} + \text{Cov}_j\left(\tau_{ij}, \frac{A_{ij}^{\eta-1}}{\int A_{ij}^{\eta-1} di}\right)\right] \qquad \text{where} \qquad \overline{\tau_j} = \int \tau_{ij} di.$$

5.1 Understanding the Losses from Risk: An Example

This subsection analyzes a simple example with 4 symmetric sectors. We leverage the tractability of this case to prove that when the productivity penalty for operating in distant sectors is higher, firms are less willing to diversify. The risk concentration has a more pronounced effect in sectors where firms have high productivity. This positive correlation between risk compensation and productivity leads to larger aggregate output losses from risk.

To see the connection between productivity patterns and the effect of risk, consider the following special case. There are four sectors, whose demand shocks are correlated as follows:

$$\Omega = \begin{bmatrix} r_0 & r_1 & r_2 & r_3 \\ r_1 & r_0 & r_3 & r_2 \\ r_2 & r_3 & r_0 & r_1 \\ r_3 & r_2 & r_1 & r_0 \end{bmatrix} ,$$

where, without loss of generality, the sectors are ordered so that $1 = r_0 > r_1 > r_2 > r_3$. A continuum of firms operate in all 4 sectors. The within-sector elasticity of substitution across firms, η, is set to 2. Each firm has a (randomly drawn)'home' sector, where its productivity is highest. Firm i, in its home sector (denoted j_i^*), has a productivity normalized to 1, i.e. $\ln A_{ij_i^*} = 0$. Its productivity in the other sectors is:

$$\ln A_{ik} = -\xi_0 - \xi_1(1 - \rho_{kj_i^*}) \qquad \forall k \neq j_i^* ,$$

where the parameters $\xi_0, \xi_1 > 0$ index the intercept and slope of the productivity penalty, i.e. the decline in productivity as a function of the distance $(1 - \rho_{kj_i^*})$ from the home sector ($\rho_{kj_i^*}$ is the correlation of sector k with the home sector of the firm). Since $\xi_1 > 0$, the penalty is higher for less correlated sectors.

Given the symmetry, the set of risk wedges is the same for every firm. With a slight abuse of notation, we denote them as $\{\tau_h\}, h = 0, 1, 2, 3$, where $\tau_{ij} = \tau_h$ if $\rho_{jj_i^*} = r_h$. In matrix form,

$$\boldsymbol{\tau} = RiskPr \cdot \sigma^2 \cdot \boldsymbol{\Omega}\, \boldsymbol{\omega} . \tag{31}$$

where $\boldsymbol{\tau} := [\tau_0 \;\; \tau_1 \;\; \tau_2 \;\; \tau_3]'$, $\boldsymbol{\omega} := [\omega_0 \;\; \omega_1 \;\; \omega_2 \;\; \omega_3]'$ with ω_h denoting the revenue share of the sector whose correlation with the home sector is r_h. For analytical tractability, we approximate the

weights $\{\omega_h\}$ using the revenue shares at the no-risk limit. When $\gamma = 0$, the revenue shares of the sector are given by

$$\omega_0 = \frac{R_{ij_i^*}}{\sum_k R_{ik}} = \frac{A_{ij_i^*}}{\sum_k A_{ik}} = \frac{1}{1 + \sum_k e^{-\xi_0 - \xi_1(1-r_k)}} \tag{32}$$

$$\omega_h = \frac{e^{-\xi_0 - \xi_1(1-r_h)}}{1 + \sum_k e^{-\xi_0 - \xi_1(1-r_k)}} \qquad h = 1, 2, 3 \, . \tag{33}$$

The loss from risk – the difference between output without risk and equilibrium output, as defined in (19) – is:

$$\mathcal{L}^{Risk} = -\ln(1 - \boldsymbol{\omega}'\boldsymbol{\tau}) \, . \tag{34}$$

The next result shows that the losses from risk are more severe when the productivity advantage of the firm's home sector is greater.

Lemma 1. (Gains to Specialization and the Output Cost of Risk)

$$\frac{d}{d\xi_0} \exp(-\mathcal{L}^{Risk}) = 2\omega_0(\boldsymbol{\omega}'\boldsymbol{\tau} - \tau_0) \;\; > \;\; 0 \, , \tag{35}$$

$$\frac{d}{d\xi_1} \exp(-\mathcal{L}^{Risk}) = 2Cov(1 - \rho_j, \tau_j) \;\; > \;\; 0 \, . \tag{36}$$

Equation (35) says that the effect of the intercept parameter ξ_0 depends on the difference between the (weighted) average τ_j and that of the home sector. Intuitively, a larger penalty shifts the firm's portfolio away from non-home sectors and towards the riskiest home sector, increasing the losses. Analogously, from equation (36), the effect of the slope parameter ξ_1 depends on the covariance between the distance from the home sector $1 - \rho_j$ and the wedge τ. When productivity declines more sharply with distance, incentives to diversify fall, shifting the composition of the firm's portfolio towards sectors closer to home (i.e. riskier). This force exerts a more negative effect on aggregate output when this shift is associated with a larger increase in τ.

6 Conclusion

This paper unifies three historically disparate perspectives on firm scope by modeling the trade-off between specialization and diversification in firm product choice. While the finance view emphasizes diversification as a tool for risk management and the management literature champions core

competency for efficiency, the macro-IO perspective traditionally focuses on a count of products or sectors. By introducing a measurable construct of product distance and revenue covariance, we unite these perspectives and provide a theory-guided framework to measure the cost of uncertainty and the resulting scope expansion. Our calibrated model reveals that risk induces firms to broaden their scope to mitigate volatility, compromising their gains to specialization, resulting in a loss of roughly 10.28% of aggregate output.

Looking forward, several extensions could be fruitful. While our current framework is static, incorporating dynamic models would allow for the study of how scope evolves over the firm lifecycle, from the rapid product switching of young startups to the entrenched breadth of superstar firms. Future research might also integrate fixed costs of expansion, particularly as information technology shifts the barriers to industry entry, to explore how these costs influence the wave of specialization seen in R&D-intensive firms (Ma, 2023). Finally, explicitly modeling the relationship between diversification and exit through bankruptcy could offer a more granular view of how risk-mitigation strategies impact firm survival and the resulting distribution of firm size and markups in the aggregate economy.

References

Acemoglu, Daron, Ufuk Akcigit, Harun Alp, Nicholas Bloom, and William Kerr, "Innovation, Reallocation, and Growth," *American Economic Review*, November 2018, *108* (11), 3450–91.

Akcigit, Ufuk, Murat Alp Celik, and Jeremy Greenwood, "Buy, Keep or Sell: Economic Growth and the Market for Ideas," Working Paper 19763, National Bureau of Economic Research December 2013.

Atkeson, Andrew and Ariel Tomás Burstein, "Innovation, Firm Dynamics, and International Trade," *Journal of Political Economy*, June 2010, *118* (3), 433–484.

Autor, David, David Dorn, Lawrence F. Katz, and Christina Patterson, "The Fall of the Labor Share and the Rise of Superstar Firms," *Quarterly Journal of Economics*, 2020, *135* (4), 1482–1541.

Ayyagari, Meghana, Asli Demirgüç-Kunt, and Vojislav Maksimovic, "The Rise of Star Firms: Intangible Capital and Competition," *The Review of Financial Studies*, 08 2023, p. hhad065.

Berlingieri, Giuseppe, Maarten De Ridder, Danial Lashkari, and Davide Rigo, "Creative Destruction through Innovation Bursts," Discussion Paper dp2095, Centre for Economic Performance, London School of Economics 2025. Revise-Resubmit, Econometrica.

Bernard, Andrew B., Stephen J. Redding, and Peter K. Schott, "Multiple-Product Firms and Product Switching," *American Economic Review*, March 2010, *100* (1), 70–97.

Boar, Corina and Virgiliu Midrigan, "Markups and Misallocation," *Working Paper*, 2023.

_ , Denis Gorea, and Virgiliu Midrigan, "The Aggregate Costs of Uninsurable Business Risk," *American Economic Review*, forthcoming. Earlier version: "Why Are Returns to Private Business Wealth So Dispersed?".

Bresnahan, Timothy, Shane Greenstein, and Rebecca Henderson, "Idiosyncratic knowledge and the diffusion of managed care," *Strategic Management Journal*, 2011.

Cai, Jie and Nan Li, "Growth Through Inter-sectoral Knowledge Linkages," *The Review of Economic Studies*, 2019, *86* (5), 1827–1866.

David, Joel M, Hugo A Hopenhayn, and Venky Venkateswaran, "Information, misallocation, and aggregate productivity," *The Quarterly Journal of Economics*, 2016, *131* (2), 943–1005.

David, Joel M., Lukas Schmid, and David Zeke, "Risk-Adjusted Capital Allocation and Misallocation," *Journal of Financial Economics*, 2022, *145* (2), 684–705.

De Ridder, Maarten, "Market Power and Innovation in the Intangible Economy," Working Paper 1931, Cambridge 2021.

Eaton, Jonathan and Samuel Kortum, "Technology, Geography, and Trade," *Econometrica*, 2002, *70* (5), 1741–1779.

Eberly, Janice C., Nicolas Crouzet, Andrea Eisfeldt, and Dimitris Papanikolaou, "Intangible Capital, Non-rivalry, and Growth," 2023. Working Paper.

Eckbo, B. Espen, ed., *Handbook of Empirical Corporate Finance Set*, Amsterdam: Elsevier, 2008.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How Much Are We Losing from Productivity Dispersion?," *Journal of Political Economy*, 2023.

Eisfeldt, Andrea L and Dimitris Papanikolaou, "The value and ownership of intangible capital," *American Economic Review*, 2014, *104* (5), 189–194.

Ekerdt, Lorenz KF and Kai-Jie Wu, *The rise of specialized firms*, US Census Bureau, Center for Economic Studies, 2024.

Hoberg, Gerard and Gordon M Phillips, "Scope, Scale and Competition: The 21st Century Firm," *The Journal of Finance*, 2025, *forthcoming* (n/a).

_ **and Gordon Phillips**, "Text-based network industries and endogenous product differentiation," *Journal of political economy*, 2016, *124* (5), 1423–1465.

Kim, J, K Lee, and H Lim, "Firm Technological Proximity and the Direction of Net Knowledge Spillovers," *Working Paper*, 2023.

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

Kumar, Saten, Yuriy Gorodnichenko, and Olivier Coibion, "The Effect of Macroeconomic Uncertainty on Firm Decisions," *Econometrica*, 2023, *91(4).*

Lewellen, Wilbur G., "A Pure Financial Rationale for the Conglomerate Merger," *Journal of Finance*, 1971, *26* (2), 521–537.

Loecker, Jan De, Jan Eeckhout, and Gabriel Unger, "The Rise of Market Power and the Macroeconomic Implications," *The Quarterly Journal of Economics*, 2020.

Ma, Yueran, Mengdi Zhang, and Kaspar Zimmermann, "Business Concentration around the World: 1900–2020," 2024. Working Paper.

Ma, Yueyuan, "Specialization in a Knowledge Economy," 2023. Working paper.

Maksimovic, Vojislav and Gordon Phillips, "Do Conglomerate Firms Allocate Resources Inefficiently Across Industries? Theory and Evidence," *Journal of Finance*, 2002, *57* (2), 721–767.

Matvos, Gregor and Amit Seru, "Resource Allocation within Firms and Financial Constraints," *The Review of Financial Studies*, 2014.

Mayer, Thierry, Marc J Melitz, and Gianmarco IP Ottaviano, "Product Mix and Firm Productivity Responses to Trade Competition," *Review of Economic Studies*, 2016.

Pellegrino, Bruno, "Product Differentiation and Oligopoly: A Network Approach," *American Economic Review*, 2025, *115* (4), 1170–1225.

Prahalad, Coimbatore Krishnarao and Gary Hamel, "The Core Competence of the Corporation," *Harvard Business Review*, 1990, *68*, 275–292.

Rao, C. Radhakrishna, "Diversity and Dissimilarity Coefficients: A Unified Approach," *Theoretical Population Biology*, 1982, *21* (1), 24–43.

Rumelt, Richard P., *Strategy, Structure, and Economic Performance*, Cambridge, MA: Harvard University Press, 1974.

Stein, Jeremy C., "Internal Capital Markets and the Competition for Corporate Resources," *Journal of Finance*, 1997, *52* (1), 111–133.

Teece, David J, Richard Rumelt, Giovanni Dosi, and Sidney Winter, "Understanding Corporate Coherence: Theory and Evidence," *Journal of Economic Behavior & Organization*, 1994.

Tella, Sebastian Di, Cedomir Malgieri, and Christopher Tonetti, "Risk Markups," Working Paper 33778, National Bureau of Economic Research 2025.

A Derivations for the Firm's Objective

This appendix shows that the firm's mean–variance objective is the second-order (small-risk) approximation of a standard expected-utility problem: a risk-averse owner/manager with Bernoulli utility u_i who consumes the firm's profit (as well as other income). A second-order expansion of expected utility in the revenue shifters (in logs) delivers the objective, with the curvature of u_i pinning down the risk-aversion coefficient γ_i.

The owner/manager of firm i has Bernoulli utility u_i that is three times continuously differentiable, strictly increasing, and strictly concave ($u_i' > 0$, $u_i'' < 0$), and consumes background (non-stochastic) wealth b_i plus the firm's profit,

$$c_i = b_i + \Pi_i, \qquad \Pi_i = \sum_j \big(P_{ij}Q_{ij} - WL_{ij}\big), \tag{37}$$

where sector-j revenue is generated by a labor choice L_{ij} made *before* shocks are realized,

$$P_{ij}Q_{ij} = (A_{ij}L_{ij})^{\alpha}\, X_j, \qquad \alpha \equiv 1 - \tfrac{1}{\eta}, \tag{38}$$

$X_j > 0$ is the stochastic sectoral revenue shifter, and WL_{ij} is deterministic conditional on the firm's information. The manager solves $\max_{\{L_{ij}\}} \mathbb{E}[u_i(c_i)]$. Write $x_j \equiv \log X_j$, $\tilde{x}_j \equiv x_j - \mathbb{E}[x_j]$, and $\Omega_{jk} \equiv \mathrm{Cov}(\log X_j, \log X_k)$, and let $R_{ij} \equiv \mathbb{E}[P_{ij}Q_{ij}] = (A_{ij}L_{ij})^{\alpha}\,\mathbb{E}[X_j]$ be expected revenue. We index the size of risk by $\varepsilon \downarrow 0$ through $\log X_j^{\varepsilon} = m_j + \varepsilon z_j$, with $z = (z_1, \ldots, z_J)'$ having finite third moments, so $\Omega^{\varepsilon} = O(\varepsilon^2)$; $o(\varepsilon^2)$ denotes remainders that hold uniformly over the shrinking choice set.

Second-order expansion

Theorem 1. *The expected-utility objective* $\max_{\{L_{ij}\}} \mathbb{E}[u_i(c_i)]$ *admits the local representation*

$$\mathbb{E}\big[u_i(c_i)\big] = u_i(\bar{c}_i) + \tfrac{1}{2}u_i''(\bar{c}_i)\sum_{j,k} R_{ij}R_{ik}\,\Omega_{jk} + o(\varepsilon^2), \tag{39}$$

where $\bar{c}_i \equiv \mathbb{E}[c_i] = b_i + \sum_j (R_{ij} - WL_{ij})$. *Equivalently, the certainty equivalent is*

$$\mathrm{CE}_i = b_i + \sum_j (R_{ij} - WL_{ij}) - \gamma_i \sum_{j,k} R_{ij}R_{ik}\,\Omega_{jk} + o(\varepsilon^2), \qquad \gamma_i \equiv -\tfrac{1}{2}\frac{u_i''(\bar{c}_i)}{u_i'(\bar{c}_i)}. \tag{40}$$

In particular, if $\Omega_{jk} = \rho_{jk}\sigma^2$, maximizing $\mathbb{E}[u_i(c_i)]$ over $\{L_{ij}\}$ is—to order $o(\varepsilon^2)$ and up to the constant b_i—equivalent to

$$\max_{\{L_{ij}\}} \sum_j (R_{ij} - WL_{ij}) - \gamma_i \sum_{j,k} R_{ij} R_{ik}\, \rho_{jk}\sigma^2, \tag{41}$$

the firm's mean–variance objective used in the main text.

Proof. Write consumption as a function of the log-shifter vector x, $c_i(x) = b_i + \sum_j (q_{ij}e^{x_j} - WL_{ij})$ with $q_{ij} \equiv (A_{ij}L_{ij})^\alpha$, and set $g(x) \equiv u_i(c_i(x))$, $r_{ij}^0 \equiv q_{ij}e^{\mathbb{E}[x_j]}$, $c_i^0 \equiv c_i(\mathbb{E}[x])$. The Hessian of g,

$$\frac{\partial^2 g}{\partial x_j \partial x_k} = u_i''(c_i)\,(q_{ij}e^{x_j})(q_{ik}e^{x_k}) + u_i'(c_i)\,(q_{ij}e^{x_j})\,\delta_{jk}, \tag{42}$$

splits into the curvature of u_i and the convexity of the exponential (level-vs-log) map. A second-order Taylor expansion of g about $\mathbb{E}[x]$, followed by taking expectations, annihilates the linear term because $\mathbb{E}[\tilde{x}_j] = 0$:

$$\mathbb{E}[u_i(c_i)] = u_i(c_i^0) + \tfrac{1}{2}u_i''(c_i^0)\sum_{j,k} r_{ij}^0 r_{ik}^0\,\Omega_{jk} + \tfrac{1}{2}u_i'(c_i^0)\sum_j r_{ij}^0\,\Omega_{jj} + o(\varepsilon^2). \tag{43}$$

The last term is the Jensen gap: since $R_{ij} = r_{ij}^0(1 + \frac{1}{2}\Omega_{jj}) + o(\varepsilon^2)$, expected consumption is $\bar{c}_i = c_i^0 + \frac{1}{2}\sum_j r_{ij}^0\,\Omega_{jj} + o(\varepsilon^2)$, the wage bills cancelling. Re-anchoring $u_i(c_i^0) = u_i(\bar{c}_i) - u_i'(\bar{c}_i)\frac{1}{2}\sum_j r_{ij}^0\,\Omega_{jj} + o(\varepsilon^2)$ in (43) cancels the two u_i' terms; replacing $r_{ij}^0 \to R_{ij}$ and $c_i^0 \to \bar{c}_i$ in the remaining $O(\varepsilon^2)$ term gives (39). Finally, $u_i(\mathrm{CE}_i) = \mathbb{E}[u_i(c_i)]$ with $\mathrm{CE}_i = \bar{c}_i + \delta_i$, $\delta_i = O(\varepsilon^2)$, gives $\delta_i = -\gamma_i \sum_{j,k} R_{ij}R_{ik}\Omega_{jk} + o(\varepsilon^2)$ by a first-order expansion, which is (40); setting $\Omega_{jk} = \rho_{jk}\sigma^2$ and dropping the constant b_i yields (41). □

Remark (mean–variance reading). To leading order the bracketed penalty is γ_i times the variance of profit, $\mathrm{Var}(\Pi_i) = \sum_{j,k} R_{ij}R_{ik}\,\Omega_{jk} + o(\varepsilon^2)$, so (40) is exactly $\mathrm{CE}_i = b_i + \mathbb{E}[\Pi_i] - \gamma_i \mathrm{Var}(\Pi_i) + o(\varepsilon^2)$.

B Derivations for Equilibrium

This appendix derives the three equilibrium relationships used in the main text: the risk-adjusted markup (9), expected sectoral output (15), and the productivity-recovery relation (20). Throughout, firms solve the operational mean–variance problem (8) obtained in Appendix A; the curvature of the owner's utility enters only through γ_i.

Firm i produces $Q_{ij} = A_j A_{ij} L_{ij}$ and, combining the inverse demand (4) with the production function, earns revenue $P_{ij}Q_{ij} = (A_{ij}L_{ij})^\alpha X_j$ as in (5), with

$$\alpha \equiv 1 - \tfrac{1}{\eta} = \tfrac{\eta-1}{\eta}, \qquad R_{ij} \equiv \mathbb{E}[P_{ij}Q_{ij}] = (A_{ij}L_{ij})^\alpha\,\mathbb{E}[X_j], \qquad R_i \equiv \sum_j R_{ij}. \tag{44}$$

Because L_{ij} is chosen before the shocks $\{A_j, B_j\}$ are realized, it is non-stochastic given the firm's information; hence revenue scales as L_{ij}^α and the marginal revenue product of labor is $\partial R_{ij}/\partial L_{ij} = \alpha\, R_{ij}/L_{ij}$. Writing $\sigma^2\rho_{jk} = \mathrm{Cov}(\ln X_j, \ln X_k)$ for the covariance of the log revenue shifters, the risk

wedge of (11)–(12) is

$$\tau_{ij} \;\equiv\; \underbrace{2\gamma_i R_i}_{RiskPr_i} \cdot \sigma^2 \underbrace{\sum_k \tfrac{R_{ik}}{R_i}\rho_{jk}}_{RiskExp_{ij}} \;=\; 2\gamma_i\,\sigma^2 \sum_k R_{ik}\rho_{jk}. \tag{45}$$

Optimal markups

Theorem 2. *The labor choice that solves* (8) *delivers the firm–sector markup*

$$\frac{R_{ij}}{WL_{ij}} = \frac{1}{1-\tau_{ij}}\,\frac{\eta}{\eta-1}, \tag{46}$$

with τ_{ij} as in (45)*; this is* (9)*.*

Proof. Differentiating (8) in L_{ij} and using $\partial R_{ij}/\partial L_{ij} = \alpha\, R_{ij}/L_{ij}$ for both the revenue and the variance term (the latter symmetric in $\rho_{jk} = \rho_{kj}$), the first-order condition is

$$\alpha\,\frac{R_{ij}}{L_{ij}} = W + 2\gamma_i\,\sigma^2\Big(\alpha\,\frac{R_{ij}}{L_{ij}}\Big)\sum_k R_{ik}\rho_{jk}. \tag{47}$$

Dividing by αW and recognizing τ_{ij} from (45),

$$\frac{R_{ij}}{WL_{ij}} = \frac{\eta}{\eta-1} + \tau_{ij}\,\frac{R_{ij}}{WL_{ij}},$$

and solving for the markup gives (46). □

Inverting (46) gives the optimal wage bill $WL_{ij} = \alpha(1-\tau_{ij})R_{ij}$. Summing over j and dividing R_i by $\sum_j WL_{ij}$ yields the firm-level markup (13), in which the undiversified-risk term is $1 - Scope_i = \sum_{j,k} \frac{R_{ij}}{R_i}\frac{R_{ik}}{R_i}\rho_{jk}$, the revenue-weighted average correlation of (14).

Equilibrium output

Define the firm-side labor aggregate and aggregate productivity

$$\widetilde{AL} \equiv \Big(\int A_{ij}^{\alpha} L_{ij}^{\alpha}\, di\Big)^{1/\alpha}, \qquad A^* \equiv \Big(\int B_j^{\frac{1}{\theta}} A_j^{\frac{\theta-1}{\theta}}\, dj\Big)^{\frac{\theta}{\theta-1}}. \tag{48}$$

Since sectors are ex-ante symmetric and L_{ij} is set before the sectoral shocks, $\widetilde{AL}$ is deterministic and common across j; with no aggregate uncertainty A^*, and hence Q, are deterministic as well.

Theorem 3. *In the symmetric equilibrium, $Q_j = A_j\widetilde{AL}$ and $Q = A^*\widetilde{AL} = W$, and expected sectoral GDP is*

$$\mathbb{E}[Y_j] = \mathbb{E}[P_jQ_j] = A^*\,\frac{\eta-1}{\eta}\Big(\int (1-\tau_{ij})^{\eta-1} A_{ij}^{\eta-1}\, di\Big)^{\frac{1}{\eta-1}}, \tag{49}$$

which is (15)*.*

Proof. The sectoral aggregator (1) with $Q_{ij} = A_jA_{ij}L_{ij}$ gives $Q_j = (\int (A_jA_{ij}L_{ij})^{\alpha} di)^{1/\alpha} = A_j\widetilde{AL}$, and substituting into (2) yields $Q = (\int B_j^{1/\theta} A_j^{(\theta-1)/\theta} dj)^{\theta/(\theta-1)}\,\widetilde{AL} = A^*\widetilde{AL}$. Household optimality

gives $W = Q$. Using $Q_j = A_j\widetilde{AL}$ and identifying the cross-sectional integral over the continuum of symmetric sectors with the expectation over their shocks,

$$\mathbb{E}[P_jA_j] = \mathbb{E}\Big[B_j^{\frac{1}{\theta}}\big(\tfrac{A_j\widetilde{AL}}{Q}\big)^{-\frac{1}{\theta}}A_j\Big] = \mathbb{E}\big[B_j^{\frac{1}{\theta}}A_j^{1-\frac{1}{\theta}}\big]\Big(\tfrac{\widetilde{AL}}{Q}\Big)^{-\frac{1}{\theta}} = (A^*)^{\frac{\theta-1}{\theta}}(A^*)^{\frac{1}{\theta}} = A^*,$$

where the last step uses $\widetilde{AL}/Q = 1/A^*$. Hence $\mathbb{E}[Y_j] = \mathbb{E}[P_jA_j]\,\widetilde{AL} = A^*\widetilde{AL}$.

To evaluate $\widetilde{AL}$, use the optimal wage bill $WL_{ij} = \alpha(1-\tau_{ij})R_{ij}$ from (46) together with the explicit revenue $R_{ij} = \mathbb{E}[P_jQ_j^{1/\eta}A_j^{1-1/\eta}]A_{ij}^{\alpha}L_{ij}^{\alpha} = \mathbb{E}[P_jA_j]\,\widetilde{AL}^{1/\eta}A_{ij}^{\alpha}L_{ij}^{\alpha}$ (the second equality from $Q_j = A_j\widetilde{AL}$). Solving for $L_{ij}^{1/\eta} = L_{ij}^{1-\alpha}$,

$$L_{ij}^{1/\eta} = \alpha\,\frac{1-\tau_{ij}}{W}\,\mathbb{E}[P_jA_j]\,\widetilde{AL}^{1/\eta}A_{ij}^{\alpha}.$$

Multiplying by $A_{ij}^{1/\eta}\widetilde{AL}^{-1/\eta}$ and substituting $\mathbb{E}[P_jA_j] = A^*$ and $W = A^*\widetilde{AL}$,

$$\Big(\tfrac{A_{ij}L_{ij}}{\widetilde{AL}}\Big)^{1/\eta} = \alpha\,\frac{A_{ij}}{\widetilde{AL}}(1-\tau_{ij}) \qquad\Longrightarrow\qquad \frac{A_{ij}L_{ij}}{\widetilde{AL}} = \Big(\alpha\,\frac{A_{ij}}{\widetilde{AL}}(1-\tau_{ij})\Big)^{\eta}. \tag{50}$$

Raising (50) to the power α and integrating over i, the left side is $\widetilde{AL}^{-\alpha}\int(A_{ij}L_{ij})^{\alpha}di = 1$ by the definition of $\widetilde{AL}$, so

$$1 = \Big(\tfrac{\eta-1}{\eta}\Big)^{\eta-1}\widetilde{AL}^{-(\eta-1)}\int A_{ij}^{\eta-1}(1-\tau_{ij})^{\eta-1}di \;\;\Longrightarrow\;\; \widetilde{AL} = \frac{\eta-1}{\eta}\Big(\int A_{ij}^{\eta-1}(1-\tau_{ij})^{\eta-1}di\Big)^{\frac{1}{\eta-1}}.$$

Combining with $\mathbb{E}[Y_j] = A^*\widetilde{AL}$ gives (49). □

Recovering productivities

Theorem 4. *Firm i's sector-j revenue share $s_{ij} \equiv P_{ij}Q_{ij}/(P_jQ_j)$ satisfies*

$$\ln A_{ij} = \frac{1}{\eta-1}\ln s_{ij} + \ln\frac{1}{1-\tau_{ij}} + \ln\frac{\eta}{\eta-1} + \ln\widetilde{AL}, \tag{51}$$

which is (20) *with the sector-wide constant* $\ln\widetilde{AL}$.

Proof. From the inverse demand (4), $s_{ij} = (Q_{ij}/Q_j)^{\alpha}$, and since $Q_{ij}/Q_j = A_{ij}L_{ij}/\widetilde{AL}$ (the common A_j cancels),

$$s_{ij} = \Big(\tfrac{A_{ij}L_{ij}}{\widetilde{AL}}\Big)^{\alpha} = \Big(\tfrac{\eta-1}{\eta}\tfrac{A_{ij}}{\widetilde{AL}}(1-\tau_{ij})\Big)^{\eta-1},$$

the second equality raising (50) to the power α. Taking logs and solving for $\ln A_{ij}$ gives (51). □

C Robustness to Atkeson-Burstein Markups

There is considerable literature on the systematic relationship between market share and markups. A potential concern is that this channel, rather than the scope channel in the model, may drive the negative relationship between measured markups and scope. To address this concern, we

replace the reduced-form size-dependent markup specification with the markup term implied by an Atkeson-Burstein nested demand structure.

Let

$$s_{ijt} \equiv \frac{R_{ijt}}{\sum_{i'} R_{i'jt}}$$

denote firm i's market share in sector j in year t. Let η denote the within-sector elasticity, which we set to $\eta = 5$, and let $\theta \leq \eta$ denote the across-sector elasticity. Define

$$\xi(\theta) = \frac{\eta - \theta}{\theta(\eta - 1)}.$$

The Atkeson-Burstein markup under no risk considerations is give by:

$$M_{ijt}(\theta) = \frac{\eta}{\eta - 1} \frac{1}{1 - \xi(\theta) s_{ijt}}.$$

When $\theta < \eta$, the markup term increases with the firm's market share within the sector.

Our model-implied firm-sector markup under the Atkeson-Burstein specification of demand:

$$\frac{R_{ijt}}{MC_{ijt} Q_{ijt}} = \frac{M_{ijt}(\theta)}{1 - \varphi(1 - \text{Scope}_{ijt})}.$$

Let $\omega_{ijt}(\varphi, \theta)$ denote the marginal-cost weight. The firm-level aggregate firm markup is given by:

$$\widehat{\mu}_{it}(\varphi, \theta) = \frac{1}{1 - \varphi(1 - \text{Scope}_{it})} \sum_j M_{ijt}(\theta) \omega_{ijt}(\varphi, \theta). \tag{52}$$

We estimate this equation using measured firm-level markups. Define residual is

$$u_{it}(\varphi, \theta) = \mu_{it}^w - \widehat{\mu}_{it}(\varphi, \theta),$$

where μ_{it}^w is the measured markup, winsorized by year at the 1st and 99th percentiles.

Our main Atkeson-Burstein exercise fixes θ and estimates only φ.

For each imposed value of θ, we use within-firm variation and the following moment condition (sample analog used for estimation) to estimate φ:

$$\frac{1}{N} \sum_{i,t} \widetilde{\text{Scope}}_{it} \widetilde{u}_{it}(\hat{\varphi}; \theta) = 0, \tag{53}$$

where tildes denote firm demeaning.

The IV estimator instruments the current measured scope with its lagged value to address concerns about contemporaneous demand shocks that move both the left- and right-hand sides of the key markup equation.

The IV moment is as follows:

$$\frac{1}{N} \sum_{i,t} \widetilde{\text{Scope}}_{i,t-5} \widetilde{u}_{it}(\hat{\varphi}_{IV}; \theta) = 0. \tag{54}$$

Everything else follows the main estimation procedure.

Table 6: Atkeson-Burstein Robustness: Calibrated Across-Sector Elasticity

Across-sector elasticity θ	$\hat{\varphi}$ Within-firm GMM	$\hat{\varphi}_{IV}$ Lagged IV GMM
4.0	0.145	0.163
	(0.017)	(0.073)
2.0	0.131	0.134
	(0.017)	(0.070)
1.5	0.114	0.109
	(0.017)	(0.068)
1.2	0.082	0.076
	(0.018)	(0.066)

Notes: The table reports one-parameter GMM estimates of φ under imposed values of the Atkeson-Burstein across-sector elasticity θ. The within-firm estimator uses equation (53). The IV estimator uses equation (54); Standard errors in parentheses are bootstrapped standard errors.

Across the calibrated values, the estimate of φ remains positive. For relatively high across-sector elasticities, the estimates are close to the baseline estimates. As θ falls and the Atkeson-Burstein market-share channel becomes stronger, the estimated price of risk declines, but the IV estimate remains positive even at $\theta = 1.2$. This indicates that the risk channel remains a significant determinant of markup even with strong market-share dependent markups.

As a complementary exercise, we also jointly estimate φ and θ. This requires a second moment to identify the market-share curvature.
The second moment uses a marginal-cost-weighted average of market shares:

$$\bar{s}_{it}(\varphi, \theta) = \sum_j \omega_{ijt}(\varphi, \theta) s_{ijt}.$$

The within-firm joint GMM estimator solves

$$\frac{1}{N} \sum_{i,t} \begin{pmatrix} \widetilde{\text{Scope}}_{it} \\ \widetilde{\bar{s}}_{it}(\varphi, \theta) \end{pmatrix} \widetilde{u}_{it}(\hat{\varphi}, \hat{\theta}) = 0, \tag{55}$$

and the IV version again instruments both objects with lagged values,

$$\frac{1}{N} \sum_{i,t} \begin{pmatrix} \widetilde{\text{Scope}}_{i,t-5} \\ \widetilde{\bar{s}}_{i,t-5}(\varphi, \theta) \end{pmatrix} \widetilde{u}_{it}(\hat{\varphi}_{IV}, \hat{\theta}_{IV}) = 0. \tag{56}$$

The joint within-firm estimate is $\hat{\varphi} = 0.147$ with $\hat{\theta} = 5.000$, so the estimate hits the upper bound $\theta = \eta$ and implies essentially no Atkeson-Burstein market-share curvature. The joint IV estimate is $\hat{\varphi}_{IV} = 0.076$ with $\hat{\theta}_{IV} = 1.204$. Standard errors reported are bootstrapped standard errors. Thus, even when the Atkeson-Burstein curvature is estimated jointly rather than imposed, the estimated price of risk remains positive, albeit with limited precision.

D LLM-Generated Shares vs. Compustat Segment Data

This appendix documents the comparison, summarized in Section 3, between the LLM-generated revenue shares and the standardized business-segment data that Compustat collects from the same filings.

Sample and segment counts. We take the union of base- and final-year observations across all firm-pairs in our data, a panel of 36,658 firm-years, and match it to the Compustat Historical Segments file by CIK and fiscal year; 27,466 firm-years have a business-segment record. Because the segment file stores multiple source vintages of the same segment, raw row counts overstate the number of segments; we count distinct segment names within a firm-year. On this basis, 15,005 firm-years (54.6%) report exactly one business segment. These single-segment firm-years are not firms declaring themselves undiversified: 34.8% of them report two or more *geographic* segments—a disclosure that satisfies the segment-reporting rules, under which firms report the units management uses to run the company, but that carries no information about the firm's composition across industries.

How often the LLM sees more than one industry. Table 7 reports, for the 15,005 single-segment firm-years, the fraction in which the LLM assigns revenue to more than one 4-digit NAICS sector, at increasing materiality thresholds for the additional sectors. The LLM reports at least one additional sector in 62% of these firm-years; requiring the second sector to account for at least 10% of revenue, the fraction is 32%. Because the segment data for these firm-years consists of a single line covering the entire firm, the additional sectors are necessarily extracted from the narrative of the filing.

Table 7: LLM sector counts for firm-years with a single Compustat business segment

Materiality threshold for second sector	Share of firm-years
Any positive revenue share	62.4%
At least 5% of revenue	43.5%
At least 10% of revenue	31.7%
At least 20% of revenue	17.7%

Notes: Sample is the 15,005 firm-years in which Compustat records exactly one business segment (distinct segment names within firm-year). Each row reports the fraction of these firm-years in which the LLM assigns positive revenue share to two or more 4-digit NAICS sectors, where the second-largest sector meets the stated threshold.

Examples. The single-segment firm-years include some of the largest firms in the sample, and the lone segment is typically a whole-firm catch-all rather than the description of a focused business. Table 8 shows four examples; in each, Compustat's one business segment covers the firm's entire revenue, while the LLM separates industries that are visible in the filing's narrative and revenue disclosures. For PG&E, the segment label itself names two industries, and the LLM's electric–gas split of 71/29 is close to the revenue split reported in the filing. For Cisco, the LLM's allocation across equipment, services, and software matches the product-versus-service revenue split disclosed on the firm's income statement (roughly 73/27). For Mondelez, the allocation across bakery products, confectionery, dairy, and beverages mirrors the product-category revenue reported in the filing. For Albertson's, the pharmacy and fuel businesses operated inside its supermarkets are real revenue lines that the segment file does not record.

Table 8: Examples: large firms with a single Compustat business segment

Firm (fiscal year)	Compustat segment (sales)	LLM sectors (4-digit NAICS: share)
PG&E Corp (2024)	"Electric & Gas Utility" ($24.4bn)	Electric power (2211): 71%; natural gas distribution (2212): 29%
Cisco Systems (2024)	"Networking Systems" ($53.8bn)	Communications equipment (3342): 54%; computer services (5415): 27%; software publishers (5112): 19%
Mondelez Intl. (2024)	"Snack Food & Beverages" ($36.4bn)	Bakery products (3118): 51%; confectionery (3113): 35%; dairy products (3115): 11%; beverages (3121): 4%
Albertson's (2006)	"Supermarket Operations" ($40.4bn)	Grocery stores (4451): 70%; pharmacies (4461): 20%; gasoline stations (4471): 5%; nonstore retail (4541): 5%

Notes: In each case the firm's single business segment in Compustat has sales equal to total firm revenue, i.e., it is a whole-firm catch-all. LLM sectors are the revenue shares assigned for the fiscal year shown; sector titles are abbreviated from the official NAICS titles (e.g., 4461 is "health and personal care stores").

Apple in fiscal 2024. Table 9 illustrates the comparison for Apple. Panel A lists every segment record Compustat holds for Apple in fiscal 2024: one business segment spanning the firm's entire $391 billion of revenue, plus five geographic segments. Apple's reportable segments are its regional operations, so the standardized data contains no industry disaggregation at all. Panel B compares the product-line revenue that Apple disclosed in the same 10-K with the LLM's output: the LLM allocates revenue across five 4-digit NAICS sectors with shares that match the disclosed product-line shares to within one percentage point, line by line.

Table 9: Apple, fiscal 2024: Compustat segment records vs. LLM-generated shares

Panel A. Compustat segment records

Segment type	Segment name	Sales ($bn)
Business segment	Mobile Communication	391.0
Geographic segment	Americas	167.0
Geographic segment	Europe	101.3
Geographic segment	Greater China	67.0
Geographic segment	Japan	25.1
Geographic segment	Rest of Asia Pacific	30.7

Panel B. 10-K product-line revenue vs. LLM output

10-K product line	$bn	Share	LLM sector (4-digit NAICS)	LLM share
iPhone	201.2	51.4%	Communications equipment mfg. (3342)	52%
Services	96.2	24.6%	Other information services (5191)	25%
Wearables, Home and Accessories	37.0	9.5%	Audio and video equipment mfg. (3343)	10%
Mac	30.0	7.7%	Computer and peripheral mfg. (3341)	8%
iPad	26.7	6.8%	Electronic component mfg. (3344)	7%
Total	391.0	100%		100%

Notes: Panel A reports all segment records in the Compustat Historical Segments file for Apple in fiscal 2024. The single business segment's sales equal the firm's total revenue, i.e., it is a whole-firm catch-all rather than a line of business. Panel B: product-line revenue is from the revenue disaggregation disclosed in Apple's fiscal 2024 Form 10-K; this detail appears outside the segment footnote and is therefore not captured in segment databases. The pairing of LLM sectors with product lines is ours, based on economic content and share magnitude. The code assigned to the iPad line (3344, electronic components) would arguably be better placed in 3341 (computers); this is an instance of the within-industry granularity differences discussed in Section 3 and does not affect the share allocation.