Intangible Intensity

Andrea L. Eisfeldt* Barney Hartman-Glaser†

Edward T. Kim‡ Ki Beom Lee§

February 2026

Abstract

We develop a text-based measure of intangible investment intensity derived from firms' 10-K filings. Our approach further classifies disclosure text into knowledge, customer, and organization capital. Firms with high intangible intensity are smaller, younger, and invest heavily in R&D and human capital. The three subcomponents map cleanly to distinct economic firm types: knowledge-intensive firms are R&D-driven with high valuations and skilled labor; customer-intensive firms are mature, profitable, and commercially oriented; and organization-intensive firms are large, asset-heavy incumbents. Managerial descriptions of expenditures thus contain informative signals about intangible investment that complement financial statements in measuring corporate capital stocks.

For helpful comments, the authors thank seminar participants at the NBER Digital Economics and AI Tutorial. *First draft: November 2025.*

*UCLA Anderson School of Management and NBER. Email: andrea.eisfeldt@anderson.ucla.edu

†UCLA Anderson School of Management. Email: barney.hartman-glaser@anderson.ucla.edu

‡Stephen M. Ross School of Business, University of Michigan. Email: etkim@umich.edu

§UCLA Anderson School of Management. Email: kibeom.lee.phd@anderson.ucla.edu

1 Introduction

Intangible assets are an important and fast-growing component of firms' capital stocks, and many studies estimate that intangibles now account for roughly one third to more than half of corporate capital (Corrado et al., 2009; Eisfeldt and Papanikolaou, 2013, 2014; Peters and Taylor, 2017; Belo et al., 2019, Ewens et al., 2020; Crouzet et al., 2022; Falato et al., 2022). Yet despite its prominence, measuring intangible investment at the firm level remains difficult, in part because much of it is expensed, unreported, or embedded in the organization. In this paper, we develop a novel text-based measure of intangible investment intensity derived from firms' 10-K filings. Using a scalable pipeline that combines large language model (LLM)-based filtering with embeddings and community detection, we construct a time-varying index of intangible intensity, θ_{int}, for all U.S. firms. Our measure captures meaningful patterns in firms' disclosure behavior and provides an integrated framework to study how intangible investment shapes firm outcomes.

Measuring intangible investment from accounting data is notoriously difficult. Selling, general, and administrative (SG&A) expenses confound operating expenses with genuine investments; research and development (R&D) and advertising are voluntary and underreported; and no line item exists for spending on organization capital. Even when data are available, it is often unclear what fraction of expenditures reflects investment rather than recurring cost. A natural alternative is to study the qualitative disclosures that managers provide when discussing their operations and cost drivers. However, extracting structured information from raw text has historically been challenging due to inconsistent reporting practices, the absence of standardized language, and the difficulty of connecting firm disclosures to economic constructs. These obstacles, together with computational limits and concerns around the interpretability of modern language models, have prevented the systematic use of disclosures to measure intangible investment at scale.

We overcome these challenges by building a text-processing framework that blends interpretability with modern natural language processing tools. We first use a lightweight LLM to isolate semantically relevant passages within Item 7 of each 10-K filing, which contains firms' discussion of costs and operating activities. We then extract a large set of economically meaningful n-grams, embed them using a sentence-transformer model, and apply filtering and clustering tools to obtain a refined semantic representation. This process yields a set of approximately 10,000 high-frequency n-grams grouped into roughly 250 coherent communities, which are then classified into intangible, non-intangible, and unknown categories. For the communities classified as intangible, we further decompose them into knowledge, brand or customer (henceforth customer)

capital. Scoring each filing involves counting the relative share of n-grams belonging to these communities. The resulting measure, θ_{int}, is a scale-agnostic intensity index: it quantifies the proportion of relevant text devoted to intangible-related activities. The method covers nearly the entire universe of public firms from 2002 to 2023 and produces stable, persistent measures with substantial variation across industries and considerable heterogeneity within industries.

Finally, we use θ_{int} to study firm characteristics. Sorting firms each year into quintiles reveals a clear economic pattern: high-θ_{int} firms are smaller, younger, less profitable, and deeply engaged in investment activities, with high R&D and SG&A intensity, substantial liquidity needs, and elevated labor costs. These firms resemble classic high-growth firms that have yet to reach sustained profitability. A major contribution of our paper is that we can disentangle this aggregate signal into its components. Sorting firms on θ_{know}, θ_{cust}, and θ_{org} reveals three economically distinct types of firms. Knowledge capital-intensive firms have the strongest growth orientation, high spending on R&D, and heavy reliance on skilled labor. Customer capital-intensive firms, in contrast, exhibit characteristics of mature, profitable firms focused on customer acquisition and commercial scale rather than innovation. Organization capital-intensive firms reflect a different dimension altogether: they are large, asset-intensive incumbents with significant operational complexity. This component has been especially difficult to measure in prior work due to the absence of an accounting proxy, and our approach provides the first scalable estimate of organization capital intensity using text.

Overall, our paper provides a new framework for understanding intangible investment intensity using textual disclosures. The measure is scalable, interpretable, and applicable whenever firms file 10-Ks, enabling new questions about firm growth, valuation, strategic positioning, and organizational capabilities. By capturing the semantic footprint of intangible activity, our approach offers a complementary lens through which to study firms in settings where accounting data are incomplete or silent.

Our research contributes to multiple strands of the literature. Because intangibles lack physical embodiment and consistent accounting treatment, measuring intangible investment is difficult. A standard approach is to accumulate SG&A expenses, as first employed by Eisfeldt and Papanikolaou (2013). Their perpetual inventory method builds on Lev and Radhakrishnan (2005), who show that including SG&A flows in productivity regressions reduces Solow (1956) residuals. Subsequent work accumulates 30% of SG&A, a fraction calibrated by Hulten and Hao (2008) from a study of six research-intensive pharmaceutical firms (see Zhang, 2013; Eisfeldt and Papanikolaou, 2014;

Peters and Taylor, 2017; Sun and Xiaolan, 2019; Falato et al., 2022). While these studies confirm that SG&A expenditures are a useful proxy for intangible investment, the fraction of expenses that are investment vs. other costs is an important outstanding question.[1]

SG&A includes items related to operations that, conceptually, should likely not be capitalized as intangible investment (Enache and Srivastava, 2018). Most importantly, there is a serious data limitation: U.S. GAAP does not require firms to separately report SG&A or its components (Markovitch et al., 2020). Although supplementing Compustat with Capital IQ data improves customer capital expense measures (He et al., 2025), there remains no systematic way to decompose SG&A. Our method improves on this along two dimensions. First, it is applicable whenever 10-Ks are available. Compared to Compustat measures of R&D and advertising expenses, which are available for 65% and 41% of firm-years in our sample, our intangible-intensity score covers 84% (and 98% after within-firm interpolation). Second, the score measures the intensity of three types of intangible capital: knowledge, customer, and organizational capital. In the absence of explicit reporting of SG&A subcomponents, our score helps shed new light on these crucial intangible assets.

Another strand of literature focuses on the use of alternative data in finance. Text data and the use of LLMs have become increasingly popular (See, for example, Gentzkow and Shapiro, 2010; Baker, Bloom and Davis, 2016; Hoberg and Phillips, 2016; Gentzkow, Kelly and Taddy, 2019; Ahci and Joos, 2024; Ottonello, Song and Sotelo, 2024; Bybee, 2025; Clayton, Coppola, Maggiori and Schreger, 2025). Most early text analyses relied on bag-of-words or keyword searches over predefined vocabularies, including applications that measure the share of discussion devoted to particular topics (Hassan et al., 2019; Kelly and Taddy, 2019). This approach is robust but can significantly limit researchers' ability to capture relevant text. In contrast, LLMs make it possible to extract richer information from text, though concerns about reproducibility and hallucinations remain. We address this by using a hybrid approach that keeps the transparency of dictionary-based methods while relying on LLMs for corpus selection and n-gram embeddings for measurement.[2]

The paper proceeds as follows. Section 2 describes our text-processing pipeline and the construction of the intangible intensity measure and its components. Section 3 documents

[1] See also Ewans, Peters, and Wang (2024), who use data on firms that have undergone mergers and acquisitions in related work.

[2] See Eisfeldt and Schubert (2025) for a survey of practical tools, pitfalls, and advice for using generative AI in finance research.

the distribution of intensity scores across firms, industries, and time. Section 4 examines firm characteristics sorted by intangible intensity. Section 5 concludes.

2 Measurement Framework and Text-Processing Pipeline

2.1 A New Measure of Intangible Investment

A central challenge in studying intangible investment is the limited informativeness of traditional accounting items. SG&A expenses are reported by the vast majority of firms, with Figure 1 showing reporting rates of roughly 90% in recent years. SG&A is valuable because it captures a broad class of expenditures that scale with firms' efforts to develop and deploy intangible assets (Eisfeldt and Papanikolaou, 2013, 2014; Eisfeldt et al., 2023).[3] At the same time, SG&A also reflects routine operating costs, making it difficult to isolate the portion that represents investment in future productive capacity. Even when SG&A rises, it is often unclear how much of that increase corresponds to activities that build intangible capital versus expenses tied to sales or short-run operational needs. Our approach addresses these limitations. While the perpetual inventory method of Eisfeldt and Papanikolaou (2013) and the subsequent literature helps smooth out one-off expenses unrelated to intangible investment, we aim to further refine existing measures by isolating text-based signals of intangible-related activities.

A second challenge concerns the specific line items most closely associated with individual intangible capital categories. Items such as R&D (a proxy for investment in knowledge capital) and advertising (customer capital) are disclosed unevenly across firms, a pattern that is clearly visible in Figure 1.[4] As of 2023, reporting rates for R&D and advertising are roughly 70% and 40%, respectively, and fewer than one third of publicly traded firms report both items. These gaps reflect pure missing data rather than zeros and severely limit the usefulness of spending-based proxies. These omissions are often not idiosyncratic; firms frequently avoid separate disclosure in ways that introduce systematic bias across industries, years, and firms. Thus, even if one could use these expense items to perfectly isolate the investment portion of SG&A, the underlying components of intangibles remain selectively reported and inconsistently measured. Our text-based measure addresses both limitations by drawing on the richer qualitative information embedded in 10-K filings.

[3] Lev and Radhakrishnan (2005) and Lev (2001) show that SG&A-funded activities such as improvements in employee incentives, internal communication systems, distribution systems, and other organizational processes contribute to the accumulation of organization capital.

[4] Investments towards organization capital are not reported as a standalone item at all.

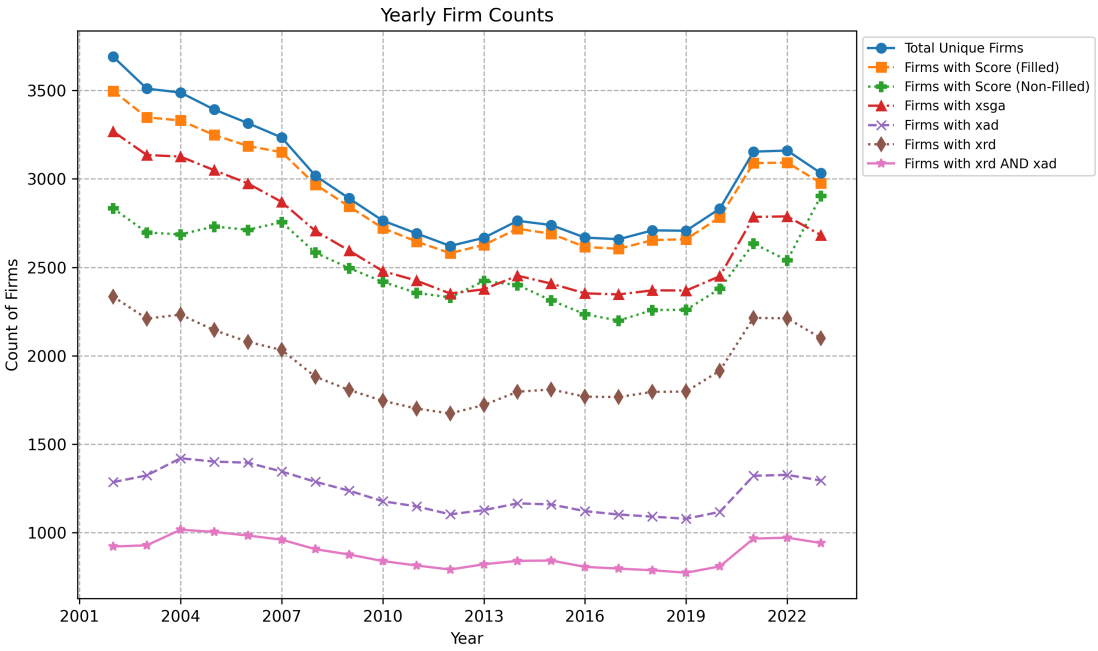


Figure 1: Coverage of Spending Variables

Notes: This figure presents the number of firms that report various spending-related line items ($xsga$, xrd, xad) in their 10-K filings.

2.2 Data Preparation

Our measure of intangible investment is constructed entirely from Form 10-K filings, which provide narrative disclosures of firms' operating activities and expenses. The sample includes all publicly traded firms from 2002 to 2023, the period over which digitized 10-K filings are readily available through the WRDS SEC Analytics Suite. We link Compustat and CRSP using standard identifiers and apply the common filters used in the accounting and finance literature to construct our sample universe. Additional details on these selection criteria and variable defitions are provided in the Appendix.

Our measure of intangible investment relies on identifying the portions of Form 10-K filings that discuss firms' investment practices. For each firm-year, we extract the Management's Discussion and Analysis (MD&A) section in Item 7, where firms typically describe SG&A expenses and other operating activities. Nearly all Form 10-K filings include an Item 7 MD&A section as standard practice, and most MD&A discussions contain at least some direct or indirect reference to SG&A. As such, MD&A provides a targeted chunk of text to analyze.

MD&A also contains discussions of capital, liquidity, and accounting policies, so relying on

the full section of text risks introducing boilerplate material unrelated to SG&A or operating expenses. To isolate the relevant text, we use a lightweight large language model, gemma3n, after first locating passages that contain SG&A or operating expense keywords. The model extracts sentences that fall into three categories of interest: *definitions*, *business drivers*, and *change drivers*. These categories mirror the way firms describe expenses: respectively, what the expense includes, the ongoing business activities it supports, and the reasons for year-over-year changes. For example, consider the following excerpt from Amazon's 2022 Form 10-K:

> *The increase in sales and marketing costs in absolute dollars in 2022, compared to the prior year, is primarily due to increased payroll and related expenses for personnel engaged in marketing and selling activities and higher marketing spend.*

In this simplified example, a single extracted sentence illustrates both the business drivers (marketing and selling activities) and the reasons for the change in SG&A from the previous year (increased payroll and spend). Business driver and change driver quotes therefore represent the variable components of SG&A, and definition quotes capture fixed or regularly incurred items that may not vary from year to year. Structuring the prompt around these three quote types allows us to recover both kinds of information. After the LLM returns the extracted sentences, we combine them into a single paragraph without duplication to form the *LLM-processed text*. We refer to the full MD&A section as the *raw text* and use the term *LLM-processed text* for this filtered output.

2.3 N-gram Dictionary and Communities

We construct an n-gram dictionary using the *LLM-processed texts* described above. Using these curated excerpts rather than the full MD&A helps prevent the dictionary from being dominated by financial boilerplate that is unrelated to SG&A. We extract bigrams and trigrams with noun–noun or noun–noun–noun structures and create a frequency-ranked "master list." Our analysis focuses on the top $N = 10{,}000$ n-grams, which balances coverage and tractability.[5]

Although n-grams offer transparency and interpretability, they capture only short spans of text and therefore lose much of the contextual information contained in full sentences. To recover this semantic structure, we embed each n-gram using a sentence-transformer model, which produces a dense vector representation that reflects the composed meaning of the phrase (for example, treating "software services" as a coherent concept rather than a bag of words). While this step mitigates

[5]Expanding the universe to 20,000 n-grams produces similar results but introduces substantially rarer phrases – the total aggregate frequency count, over all firms and years, of the least common n-gram falls from 7 to 4.

the inherent limits of n-grams, pre-trained embeddings on their own can blur distinctions between closely related spending descriptions. For example, an embedding model may place "software development services," which reflects investment in knowledge capital, very close to "software license maintenance," which is a routine operating expense.

To address these issues, we refine the raw embeddings through a sequence of post-processing steps that correct known distortions in transformer-based vector spaces and improve downstream clustering. We begin by applying Principal Component Analysis (PCA) to reduce dimensionality and remove low-variance directions that often reflect noise rather than meaningful semantic variation. Whitening further standardizes the embedding vectors by scaling all principal components equally, which corrects for the directional bias (anisotropy) that is common in pre-trained embedding models. Finally, we normalize each vector to unit length, ensuring that cosine similarity provides a stable measure of semantic relatedness. Together, these steps substantially denoise the embedding space and sharpen the separation between intangible-related and non-intangible phrases.[6]

The resulting embeddings are clustered using K-Means, which generates thousands of fine-grained micro-clusters. We treat the micro-cluster centroids as nodes in a graph and apply the Leiden community detection algorithm to identify coherent, higher-level communities of semantically related n-grams. This yields 247 communities. Each community is then hand-classified into one of five categories: knowledge, customer, organization, non-intangible, or unknown. The first three categories collectively define the set of intangible-related n-grams.[7] These communities form the backbone of our dictionary. Each n-gram is mapped to exactly one community and therefore to one of the five categories. The resulting classification system allows us to identify whether a phrase is related to intangible investment, and if so, which specific type of intangible capital it reflects.

To assess the validity of these communities, we conduct a Monte Carlo simulation that benchmarks each community's semantic cohesion against a statistically generated null. Using an independent transformer model to embed the full universe of $N = 10{,}000$ n-grams, we repeatedly draw random groups of n-grams that match the size distribution of our observed communities and compute their average cosine similarity. Figure 2 compares these null distributions to the cohesion

[6]Without these refinements, contextual embeddings can overweight boilerplate language and place economically distinct phrases too close together. The adjustments mitigate these distortions and improve the reliability of similarity-based clustering.

[7]An example set of communities and our classifications is reproduced in Appendix Table B.1.

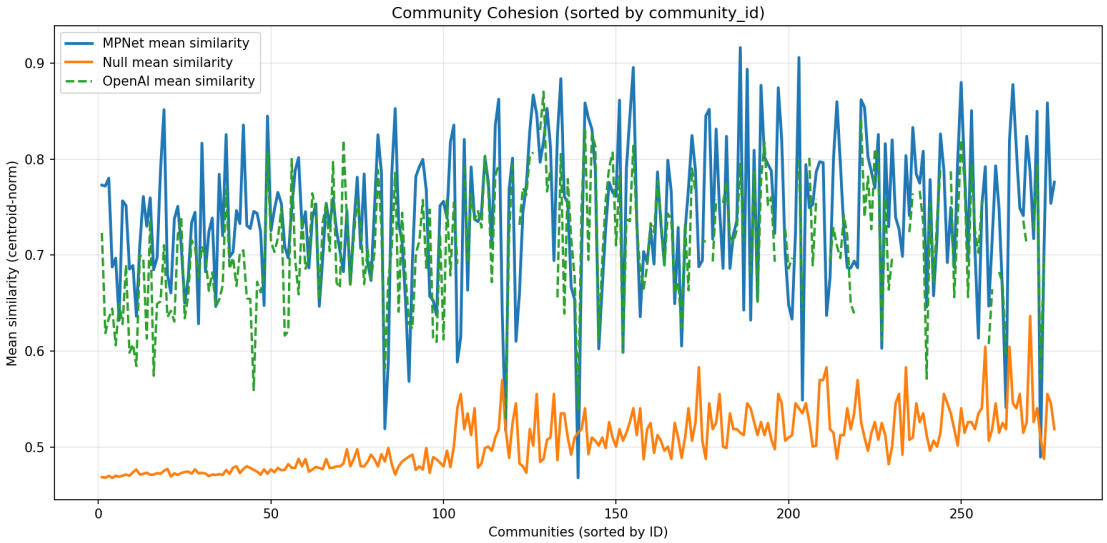


Figure 2: Community Cohesion Benchmarking

Notes: This figure benchmarks the cohesion of identified communities against a null distribution generated from random n-gram groupings embedded with an independent model.

of our actual communities. Across all community sizes, the observed cohesion substantially exceeds the Monte Carlo baseline, providing strong evidence that the communities identified by our pipeline capture meaningful semantic structure rather than spurious clustering artifacts.

In summary, the n-gram dictionary and community-generation pipeline maps tens of thousands of text fragments into a manageable set of semantic categories aligned with knowledge, customer, and organizational capital. This forms the foundation for our text-based measure of intangible investment, which we discuss next.

2.4 N-gram Scoring and Intangible Scores

Our goal in this step is to translate the dictionary of n-grams into document-level measures of intangible intensity. The starting point is the whitelist W, which contains the top $N = 10{,}000$ n-grams that have been assigned to a community. Each community is then assigned a category and (if applicable) intangibles subcategory. Specifically, each n-gram belongs to exactly one community, one of the three main categories (intangible, non-intangible, unknown), and if it is intangible, to one of the three subcategories (knowledge, customer, organization).

To construct scores, we extract n-grams from the *raw text* from MD&A rather than the

LLM-processed text. The latter is well suited for building the dictionary because it isolates SG&A-related content, but using it to score documents would artificially inflate the share of intangible-related n-grams.[8] Working with the full MD&A gives a more accurate representation of how much semantic "attention" a firm allocates to intangible-related activities relative to other topics. At the same time, restricting attention to the whitelist W ensures comparability across firms and avoids contamination from boilerplate or idiosyncratic language. This approach trades some coverage for consistency, but it provides a stable basis for measurement.

Figure 3 illustrates the distribution of n-gram coverage in the raw texts. The whitelist W captures a substantial and consistent fraction of each firm's MD&A narrative, with coverage rising smoothly in the size of the n-gram universe. We adopt $N = 10{,}000$ as our baseline (robust to larger samples), which offers a good balance between coverage and tractability.

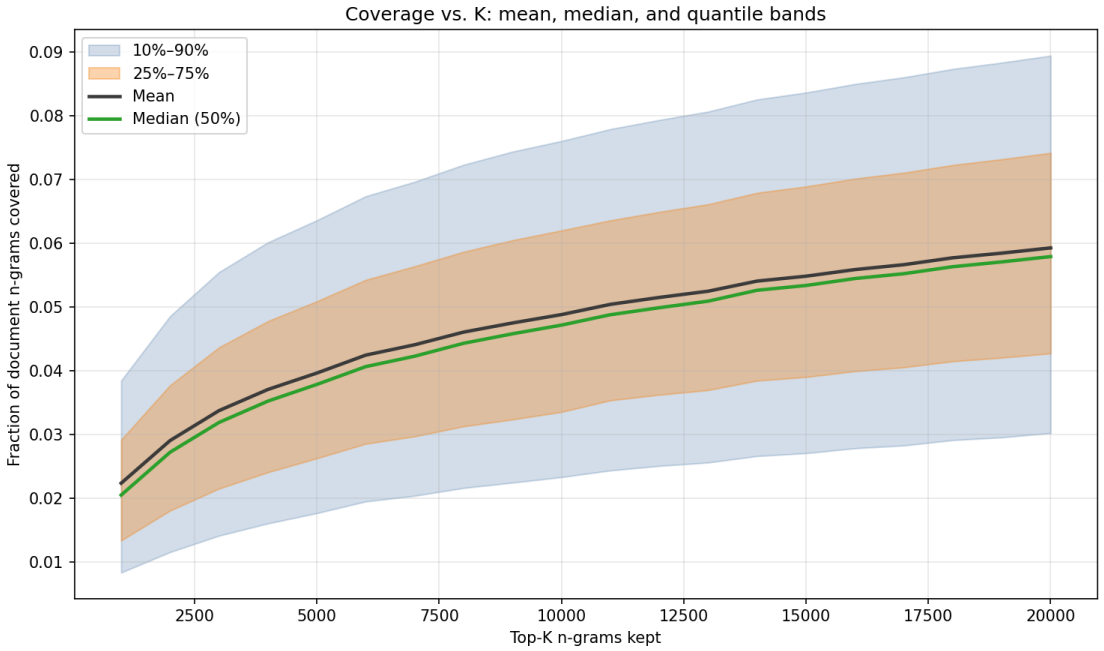


Figure 3: Whitelist Length and N-grams Coverage

Notes: This figure reports the fraction of a given document's n-grams covered as the number of whitelist n-grams varies.

The procedure for converting n-grams into document-level scores is formalized below.

[8] In extreme cases, the LLM-processed text for a firm-year may consist of only a few SG&A-focused sentences, which would distort any measure expressed as a fraction of total n-grams.

1. **Data:** For each document d_i, extract a set of n-grams $S(d_i)$ and $A(d_i) := S(d_i) \cap W$. For each n-gram $n \in A(d_i)$, let $v_n \in \mathbb{R}^E$ be a normalized embedding of n. We define the observed count of the n-gram n, $cnt_d(n) \in \mathbb{N}$.

2. **Counting Measure:** For each document d_i, we define category (c) and subcategory (s) level count measures,

$$\text{cat_cnt_sum}_d(c) = \sum_{n \in A(d_i): cat(n)=c} \text{cnt}_d(n) \tag{1}$$

$$\text{sub_cnt_sum}_d(s) = \sum_{n \in A(d_i): sub(n)=s} \text{cnt}_d(n) \tag{2}$$

   For each category c, we sum the total number of appearances of n-grams that belong to that category, and we do the same for each subcategory. By definition,

$$\text{cat_cnt_sum}_d(\text{intangible}) = \sum_{s \in Sub} \text{sub_cnt_sum}_d(s) \tag{3}$$

   Similarly, we can define an analogous count measure at the community level (h), which is useful for industry-level analyses.

$$\text{comm_cnt_sum}_d(h) = \sum_{n \in A(d_i): comm(n)=h} \text{cnt}_d(n) \tag{4}$$

3. **Probability (Score):** For each document d_i, we define an implied probability distribution over categories and subcategories. For category (c) and subcategory (s)

$$\theta_d(c) = \frac{\text{cat_cnt_sum}_d(c)}{\sum_{c' \in Cat} \text{cat_cnt_sum}_d(c')}, \tag{5}$$

$$\theta_d(s) = \frac{\text{sub_cnt_sum}_d(s)}{\sum_{s' \in Sub} \text{sub_cnt_sum}_d(s')}. \tag{6}$$

   And by definition,

$$\theta_d(\text{int}) + \theta_d(\text{non-intangible}) + \theta_d(\text{unknown}) = 1 \tag{7}$$

$$\theta_d(\text{know}) + \theta_d(\text{customer}) + \theta_d(\text{org}) = 1 \tag{8}$$

To simplify notation, we write firm-level $\theta(\text{int})$ as θ_{int}, $\theta(\text{know})$ as θ_{know}, $\theta(\text{customer})$ as θ_{cust},

and θ(org) as θ_{org}. These scores can be interpreted in more than one way. At a basic level, they function as semantic attention measures: for example, θ_{int} reflects the share of the MD&A narrative devoted to activities or expenses related to intangible investment. The subcategory measures offer finer distinctions. A high value of θ_{know} indicates discussion concentrated on knowledge-related activities, while elevated θ_{cust} or θ_{org} values point to an emphasis on customer-facing capabilities or organizational processes.[9]

More important for our purposes, we adopt a second interpretation and view θ as a measure of *investment intensity*. By investment intensity, we mean that θ serves as a proxy for the fraction of the firm's SG&A or operating expenses that represent intangible investment. This interpretation guides our empirical analysis, and we return to it throughout the paper.

3 Intangible Intensity from an Embeddings Model

3.1 Intangible Intensity Measure

We begin by describing the basic properties of our intangible intensity score, θ_{int}, and its coverage in our sample. Table 1 reports the number of firm-year observations in the universe of Compustat–CRSP matched 10-K filings from 2002 to 2023, and the share for which we are able to compute a nonmissing θ_{int}. We construct θ_{int} for 55,145 firm-year observations, which corresponds to roughly eighty six percent of the full sample of firm-year observations. To limit the impact of missing values on panel analyses, we linearly interpolate θ_{int} at the firm level.[10] This procedure preserves the time-series structure of the filing language and raises coverage to one hundred percent of the sample.

Table 2 reports the distribution of θ_{int}, its three subcomponents, and the non-intangible share (θ_{nint}). The average value of θ_{int} in the full sample is approximately 0.19 and the corresponding average for the non-intangible share is 0.48, confirming that the majority of the discussion in the relevant sections of 10-K filings is devoted to activities that we classify as non-intangible. The interquartile range for θ_{int} runs from roughly 0.07 to 0.29, indicating substantial dispersion across firms. Decomposing θ_{int} into its subcategories reveals that customer capital and organization capital account for larger fractions of the score than knowledge capital, which is consistent with the structure of SG&A expenses and their close connection to sales, marketing, and operations.

[9] A code repository replicating the intangible intensity pipeline is available at `https://github.com/bhglaser/semantic-theme-scoring`.

[10] Missing values for θ_{int} can arise from incomplete 10-K information, unparsed Item 7 text, or other data issues.

Table 1 Summary of Firm-Year Observations and Score Counts

Statistic	**Total**	**Average (per year)**
Total Firm-Year Observations	64,028	2,910
Non-Empty Measures θ	55,145	2,507
Non-Empty Interpolated Measures θ	64,028	2,910

Table 2 Distribution of θ Scores and Subcomponents

Type	Mean	SD	P25	P50	P75	Min	Max
θ_{int}	0.194	0.156	0.071	0.143	0.291	0.000	1.000
θ_{nint}	0.483	0.135	0.392	0.489	0.576	0.000	1.000
θ_{cust}	0.342	0.257	0.143	0.308	0.500	0.000	1.000
θ_{know}	0.263	0.277	0.000	0.182	0.480	0.000	1.000
θ_{org}	0.367	0.283	0.143	0.300	0.560	0.000	1.000

Figure 4 summarizes the evolution of θ_{int} over time. The mean and median values of the score are stable, fluctuating within a narrow band around 0.18 to 0.20 for most of the sample period. The high coverage rate and the persistent dispersion across firms are important for two reasons. First, high coverage alleviates the central limitation of traditional spending-based measures of intangible investment. Second, the wide interquartile range suggests meaningful heterogeneity in the way firms discuss intangible investment in their filings, which provides the variation needed for the cross-sectional tests that follow.

In the bottom panel of Figure 4, the ratio of SG&A to total assets is similarly stable at around 40% throughout the sample. While the units are not directly comparable to our text-based measure, the comparison is informative: in several periods, notably during the 2008 financial crisis and again in 2020 and 2021, the two series diverge; SG&A spending intensity rises when θ_{int} remains flat or declines. This implies that the information conveyed in 10-K filings does not mechanically mirror reported spending. Our intensity measure, while likely correlated with underlying expenditures, appears to capture an additional dimension of firms' emphasis that may reflect strategic communication, forward-looking guidance, or qualitative information not embedded in accounting flows. Consequently, it remains an empirical question whether θ_{int} provides incremental insight into economic outcomes beyond what can be learned from conventional spending-based proxies.

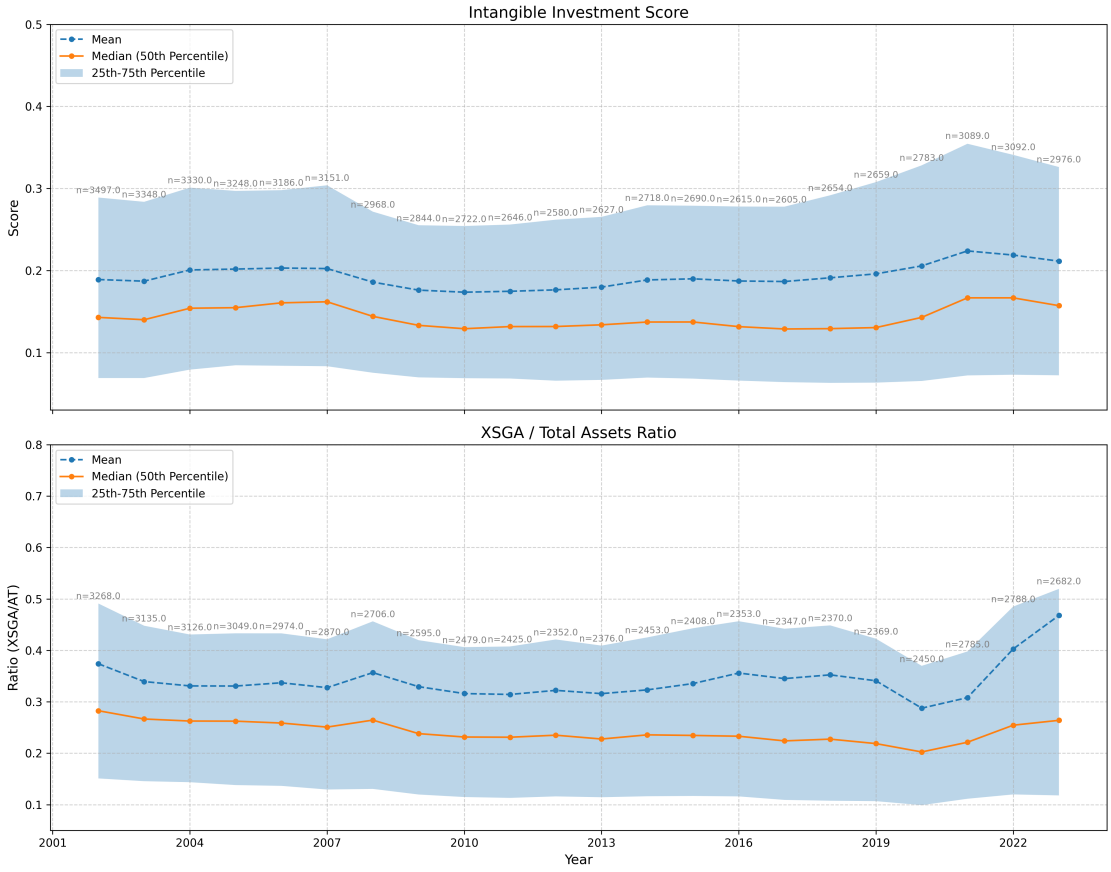


Figure 4: Distribution of Intangible Intensity Over Time

Notes: This figure reports the mean, median, and interquartile range of θ_{it} for all firm-years in the sample.

3.2 Industry Dynamics of Intensity

The wide distribution of intensity scores in Figure 4 masks substantial heterogeneity across sectors. We expect industries that rely heavily on scientific knowledge or product development to display higher levels of intangible intensity, while industries focused on physical capital to exhibit lower values.

We document these facts in Figure 5, by plotting intensity measures across the Fama-French 17 industries.[11] The intangible intensity of 0.35 to 0.40 for the "Drugs, Soap, Perfumes, and Tobacco" industry is roughly twice the aggregate average (top-left panel).[12] In contrast, industries with

[11]We drop the "Financials" industry from our sample. In unreported analysis, we verify that using NAICS industry classifications yields similar results.

[12]Firms in this industry likely exhibit substantial within-industry dispersion in terms of business segment and

the lowest average intangible intensity include "Oil and Petroleum Products," "Utilities," and "Steel Works," among others, consistent with their dependence on tangible capital and established production processes. At the other end of the spectrum, "Machinery and Business Equipment" and the residual "Other" industry, which contains many high technology and business services firms, display elevated and persistent levels of intangible intensity. Taken together, these patterns indicate that industry-level intensity measures are highly persistent and closely aligned with across-industry differences in technological and organizational demands.

We also study the decomposition of these patterns by intangible capital component, illustrating how each component contributes to industry-level intensity. The "Drugs, Soap, Perfumes, and Tobacco" industry stands out for its disproportionately high knowledge capital intensity, which is significantly higher than that of any other industry and likely driven by the presence of biotechnology and life-sciences firms. Several other industries display notable patterns as well; the knowledge capital share is almost zero for "Retail Stores," has been steadily declining among "Machinery and Business Equipment" firms, and shows a modest upward trend within the residual "Other" industry, consistent with the increasing importance of software and digital platforms. Both in levels and in shares, customer capital is most prominent in the "Food" and "Textiles, Apparel, and Footwear" industries, reflecting the centrality of consumer-facing differentiation. In contrast, organization capital plays an outsized role in "Mining and Minerals," "Oil and Petroleum Products," and "Transportation," although these patterns likely reflect the fact that many supply chain and logistics-related activities are classified as organization capital. Organization capital also features prominently among "Retail Stores," where complementarities with human capital and operational scale are particularly salient. These component-level differences reinforce the idea that intangible intensity captures distinct forms of firm capabilities and that the relative importance of each type varies systematically across industries.

Appendix Table B.2 reports the industry-level summary statistics for θ_{int} and its subcategories. We confirm both the large across-industry differences and nontrivial within-industry dispersion. In particular, the interquartile ranges are wide in nearly all industries, which underscores the importance of exploiting firm-specific variation in the empirical analysis that follows.

innovation strategies.

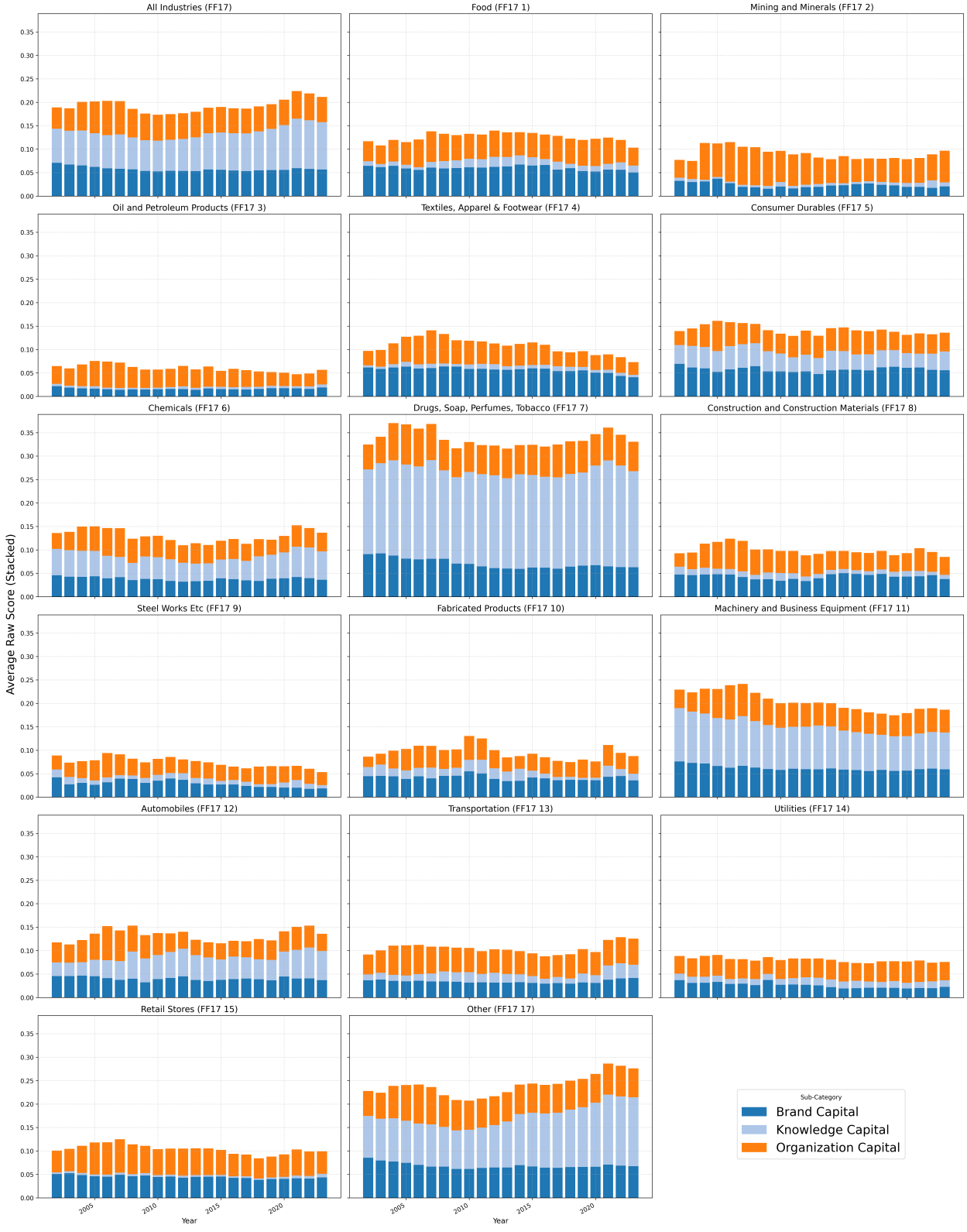


Figure 5: Intangible Intensity and Subcomponents Across Industries

Notes: This figure reports the average θ_{int} and its subcomponents for each Fama-French 17 industry excluding financials.

4 What Intangible Intensity Reveals About Firms

4.1 Firm Characteristics Across Intangible Intensity Quintiles

In this subsection, we apply θ as a sorting variable to examine the information embedded in firms' narrative disclosures. For each year, we sort firms into quintiles based on their θ values (Q1 as the lowest and Q5 as the highest) and study average firm characteristics within each group. Because θ varies substantially across industries, all sorting is conducted within industry–year cells to ensure that results are not driven by cross-industry differences. We use the Fama–French 17 industry classification as our baseline.

4.1.1 Total Intangible Intensity, θ_{int}

We begin by examining average firm characteristics sorted on θ_{int}. Table 3 shows a highly monotonic pattern across most variables, indicating that the text-based measure captures a distinct underlying economic signal. To start, high θ_{int} firms (Q5) are considerably smaller than low θ_{int} firms (Q1) across measures of market capitalization, book equity, and total assets. Despite their smaller size, these high intensity firms command sharply higher valuations: the average Tobin's Q rises from 1.34 for Q1 to 3.15 for Q5. This valuation premium persists even when adjusting the book value of equity to include the stock of intangible capital; the adjusted book-to-market ratio is 2.45 for Q1 relative to 1.76 for Q5.

While valuation ratios indicate that market expectations of future free cash flows are high, profitability (defined as the ratio of income before extraordinary items to book value of assets) exhibits the opposite relationship across quintiles. Profitability deteriorates from 0.059% to –0.258% as we move from Q1 to Q5, consistent with high-intensity firms entering a high-growth, high-expenditure phase. This pattern reflects the well-known asymmetry in accounting treatment: tangible investment is capitalized, while intangible investment is expensed, so book assets and earnings both understate the true scale of investment for high-intangible firms. This fact also aligns with their escalating intangible and operational expenditure intensity: the R&D expense to sales ratio rises from 0.025 to 12.1, and the SG&A expense to sales ratio increases from 0.213 to 2.64. Investment activity also tends to expand along both intangible (0.244 for Q1 vs. 0.296 for Q5) and physical dimensions (0.092 for Q1 vs. 0.152 for Q5).

Balance sheet and input cost patterns reinforce this interpretation. Leverage and asset tangibility decline with intangible intensity, consistent with lower collateralizability of assets.

Table 3 Characteristics by Intangible Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Market Equity	6.61	6.55	6.32	6.07	5.66
Log Book Equity	6.02	5.87	5.52	5.14	4.67
Log Total Assets	6.94	6.71	6.22	5.73	5.16
Log Intangible Capital	6.3	6.34	6.1	5.78	5.24
Log Capital-Labor Ratio	4.85	4.79	4.73	4.7	4.73
Book-to-Market	0.773	0.701	0.625	0.566	0.535
Tobin's Q	1.34	1.56	2.03	2.6	3.15
Intangible-Adjusted Book-to-Market	2.45	2.45	2.27	2.05	1.76
Total Intangible Capital / Assets	0.935	1.16	1.41	1.59	1.43
Sales / Total Intangible Capital	4.43	2.55	1.95	1.33	1.13
Profitability	0.0733	0.0589	0.00847	-0.0804	-0.232
Asset Tangibility	0.57	0.534	0.459	0.401	0.337
Solow Residual	-0.0287	0.0602	0.118	0.0684	-0.219
Intangible Capital Investment Rate	0.267	0.276	0.287	0.313	0.356
Physical Capital Investment Rate	0.102	0.117	0.138	0.164	0.222
Firm Age	24.5	22.5	20.2	16.6	12.6
Sales / Assets	1.1	1.08	0.975	0.82	0.633
Cash / Assets	0.106	0.139	0.226	0.337	0.475
R&D Expense / Assets	0.0176	0.0405	0.0895	0.155	0.255
Advertising Expense / Assets	0.0283	0.0288	0.0292	0.0315	0.0415
Debt / Assets	0.313	0.277	0.221	0.185	0.148
SG&A Expense / Assets	0.219	0.271	0.343	0.425	0.477
Employees / Assets	0.007	0.00614	0.00498	0.00398	0.00301
SG&A Expense / Sales	0.213	0.294	0.614	1.04	2.64
R&D Expense / Sales	0.0248	0.0615	0.577	3.17	12.1
Advertising Expense / Sales	0.0274	0.0295	0.0339	0.0388	0.0587
Revenue per Employee	456	447	441	404	334
Labor Expense per Employee	68.3	71.1	79.8	102	129
Executive Compensation / Assets	2.65	3.17	4.04	5.33	5.88
θ_{int}	0.0417	0.105	0.18	0.266	0.398
θ_{know}	0.0907	0.178	0.275	0.351	0.456
θ_{cust}	0.377	0.375	0.346	0.325	0.264
θ_{org}	0.402	0.445	0.377	0.322	0.278

Notes: This table presents the equal-weighted average of firm-level statistics for portfolios sorted on θ_{int}. All sorts are done within the Fama–French 17 industry classification. The sample period is from 2002 to 2023.

Highθ_{int} firms also rely more heavily on liquidity (measured as cash-to-assets) to support sustained "cash burn." Labor inputs show an even sharper shift. Labor expense per employee rises from 68.3 for Q1 to 129 for Q5, and executive compensation relative to assets more than doubles from 2.65 to 5.88. At the same time, employees relative to assets fall steadily, so high–θ_{int} firms are not reducing

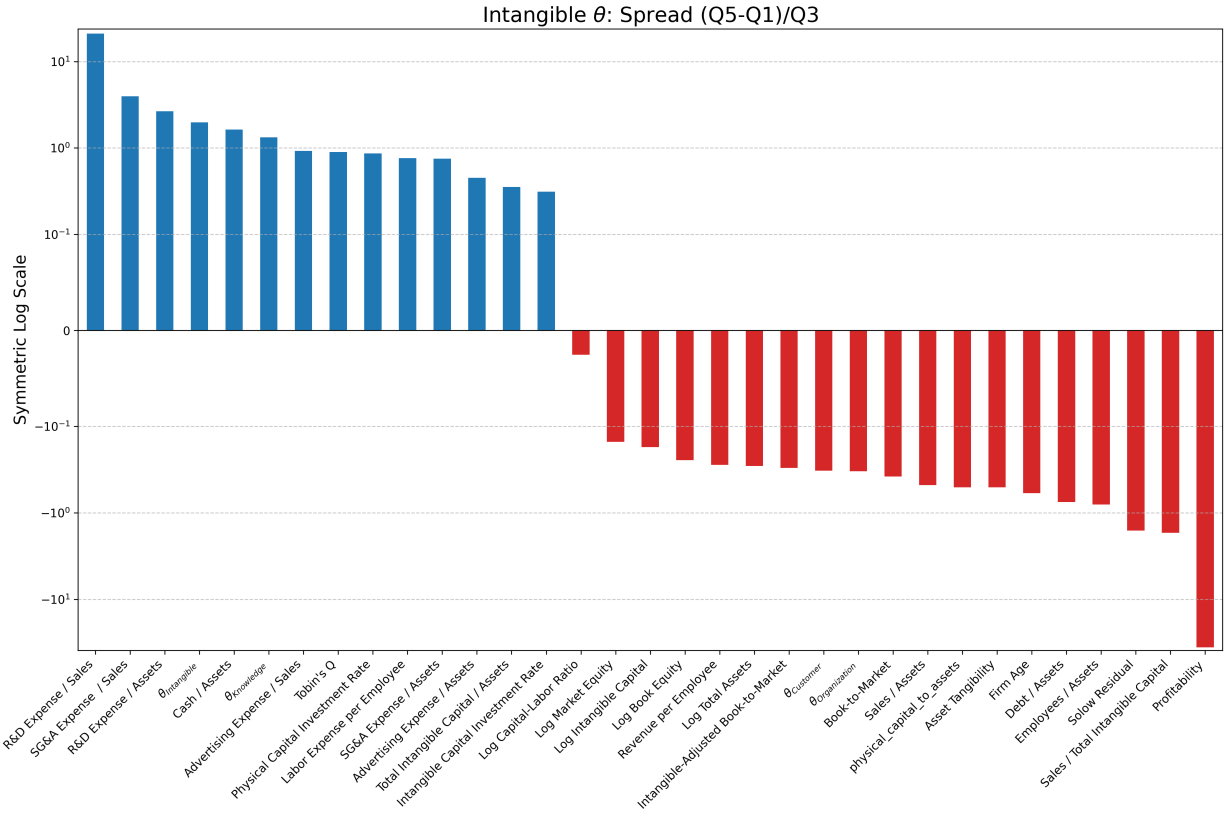


Figure 6: Q5-Q1 firm characteristics by θ

Notes: This figure reports the average difference of firm characteristics between Q5 and Q1 sorted firms by θ_{int}.

labor costs but instead allocating resources toward a smaller, more human-capital-intensive workforce.

Figure 6 provides a visual summary of these patterns by plotting the normalized difference in average characteristics between Q5 and Q1 firms. The figure reinforces the strong monotonic shifts observed in the table, particularly for valuation ratios, spending intensity, and labor-related measures. It also highlights a few secondary patterns, such as the decline in sales-to-assets and revenue-per-employee, which are consistent with firms operating at smaller scale during high-growth phases. Overall, the figure underscores how sharply high–θ_{int} firms differ from their low-intensity counterparts across a broad set of fundamentals.

Taken together, these patterns shed light on the dynamics of the firm lifecycle and how firms allocate resources as they transition from early-stage growth to more established operations. High–θ_{int} firms are smaller, fast-growing, pre-profitable firms that devote substantial resources to innovation, talent, and expansion. Perhaps the most notable result is that θ_{int}, which is

constructed entirely from text without using accounting data, identifies these aggressive spending and investment behaviors with remarkable clarity.

4.1.2 Sorts using Intangible Intensity Components

We next repeat the sorting exercise using the three subcomponents of intangible intensity, θ_{know}, θ_{cust}, and θ_{org}. This allows us to disentangle how knowledge, customer, and organization capital relate to firm characteristics.

Knowledge capital intensity (θ_{know}). Table 5 shows that sorting on θ_{know} produces patterns that closely mirror those for overall θ_{int}. High θ_{know} (Q5) firms are smaller than Q1 firms but command higher valuations (Tobin's Q of 2.68 vs. 1.4), indicating that the market places a premium on knowledge intensity. Profitability deteriorates from 0.049 to -0.198 with intensity, consistent with firms entering an intensive investment phase. Intangible and operational spending increases sharply across quintiles: the R&D to sales ratio rises from 0.0256 to 12.3, and the SG&A to sales ratio increases from 0.255 to 2.28. High θ_{know} firms also hold more liquidity, as the cash to assets ratio increases from 0.121 to 0.403, and they use less leverage, with the debt to assets ratio falling from 0.275 to 0.186. Labor inputs appear to shift toward high-skilled talent (as indicated by employees counts falling), and labor expense per employee rises with higher knowledge intensity. The top panel of Figure 7 reinforces these patterns, particularly the large Q5–Q1 spreads in R&D intensity, liquidity, and human-capital measures.

Interestingly, firms that rank in the top θ_{know} quintile within their industries also tend to have lower θ_{cust} and θ_{org}, which reinforces that knowledge capital is the primary force behind the aggregate θ_{int} signal rather than broad strength across all components.

Customer capital intensity (θ_{cust}). Table 6 provides average characteristics for five portfolios sorted on θ_{cust}. We find that θ_{cust} produces a firm profile that contrasts sharply with the patterns observed for knowledge intensity. High θ_{cust} firms resemble established, commercially focused firms rather than early-stage growth firms, as their market valuations are lower as measured by Tobin's Q, indicating.

In contrast, profitability is highest for Q5 firms, implying that customer-intensive firms generate steady earnings from an established customer base. The spending mix also differs sharply from θ_{know} sorts; the R&D to sales ratio is lowest for Q5 firms, while their advertising to sales ratio is the highest. This pattern suggests a substitution rather than complementarity between knowledge

capital and customer capital. This provides independent verification that our text-based intensity measure is highly correlated with the component of SG&A expenditure that firms do report. Further, the SG&A to sales ratio is lowest for high θ_{cust} firms, indicating a general shift away from innovation-oriented expenditures. Unlike high θ_{know} firms, High θ_{cust} firms hold less liquidity and rely more on debt.

Patterns in labor inputs across portfolios also diverge from those observed for knowledge capital. Labor expense per employee declines as θ_{cust} rises while the number of employees increase, suggesting a broader workforce with lower average cost. The middle panel of Figure 7 reinforces these patterns, highlighting strong negative Q5–Q1 spreads in R&D intensity and labor cost per employee, alongside positive spreads in advertising intensity and leverage.

In summary, θ_{cust} identifies profitable, commercially oriented firms that appear to allocate resources toward maintaining and expanding market position rather than pursuing frontier innovation. The divergence between standard and intangible-adjusted book-to-market ratios seems to indicate that the market effectively discounts customer capital, treating it more as an ongoing cost of doing business than as a source of high-growth options.

Organization capital intensity (θ_{org}). Table 7 produces a firm profile that differs markedly from the patterns associated with θ_{int} and θ_{know}. High θ_{org} firms are larger and exhibit more modest valuations: Tobin's Q declines from 2.05 for Q1 to 1.73 for Q5, while both the book-to-market ratio and the intangible-adjusted book-to-market ratio remain relatively flat. Profitability improves from -0.063 to 0.0177 as θ_{org} increases.

The composition of investment and financing reflects an established, asset-intensive profile rather than a growth-oriented one. This may be because organization capital complements tangible assets, such as process improvements that increase the productivity of a firm's physical capital. Specifically, the R&D to assets ratio falls from 0.142 for Q1 to 0.0455 for Q5, and the SG&A to sales ratio declines from 0.625 to 0.392. Asset tangibility and leverage rise with intensity: the physical capital to assets ratio increases from 0.457 to 0.525, and the debt to assets ratio increases from 0.211 to 0.270. Liquidity moves in the opposite direction, with the cash to assets ratio falling from 0.269 to 0.158. Labor patterns also differ from those seen for knowledge or customer capital. Employees to assets increase from 0.005 to 0.007, while labor expense per employee declines slightly from 87.9 to 81.3, indicating a larger workforce geared toward operational scale rather than specialized expertise. The bottom panel of Figure 7 highlights these patterns, showing

positive Q5–Q1 spreads in size, tangibility, and leverage and negative spreads in R&D intensity and liquidity.

Overall, θ_{org} captures the organizational complexity of established firms rather than a distinct high-growth regime. This pattern is consistent with the inherent difficulty of isolating organization capital from standard accounting data. Our measure offers, to our knowledge, the first comprehensive, text-based estimate of organization capital intensity in a setting where no conventional accounting proxy exists.

To sum, these patterns across subcomponents shed light on how intangible intensity maps to firms' lifecycle dynamics and investment behavior. Table 4 shows that the four θ measures move in systematically different ways. θ_{know} is strongly positively correlated with θ_{int}, whereas θ_{cust} and θ_{org} are negatively correlated with it and with each other. This structure indicates that firms tend to focus on one intangible strategy rather than pursuing all dimensions simultaneously. Appendix Figure B.2 summarizes the average characteristics across quintiles for θ_{int}, θ_{know}, θ_{cust}, and θ_{org} for a set of key variables. The figure displays the same strong monotonic relationships observed in the tables and confirms that our text-based measure – constructed entirely from narrative filings and free of numerical inputs – embeds economically meaningful information about firms profiles.

Table 4 Correlation Matrix of Intensity θ

	θ_{int}	θ_{know}	θ_{cust}	θ_{org}
θ_{int}	100.0%	66.7%	-19.9%	-30.8%
θ_{know}	66.7%	100.0%	-34.8%	-51.4%
θ_{cust}	-19.9%	-34.8%	100.0%	-43.9%
θ_{org}	-30.8%	-51.4%	-43.9%	100.0%

4.1.3 Evidence from Large Firms

As a simple external check, Table B.4 reports our intensity rankings for a set of large firms in 2021. Although these firms are at the very top of the market cap distribution and do not necessarily represent the small, high-growth firms that drive much of the cross-sectional variation, the patterns around θ_{int} and its components are broadly consistent with their business models and accounting data.

In particular, knowledge scores are highest for technology and platform firms such as Microsoft,

Alphabet, Nvidia, Tesla, and Meta, which also exhibit elevated R&D to sales ratios. Customer scores are highest for consumer-facing franchises such as Procter & Gamble, Coca-Cola, and PepsiCo, where advertising to sales ratios are large and sustained. Organization scores are relatively high for operationally intensive retailers such as Walmart and Home Depot, which have low R&D and modest customer scores but rely heavily on logistics and scale. The within-industry quintile placements also follow intuitive patterns with only a few exceptions.[13]

These cases show that the θ components align with familiar accounts of firms' competitive strengths. It also underscores the importance of accurate industry classifications, an independent question beyond the scope of this paper.

[13]For instance, Apple falls into the lowest θ_{int} quintile within its industry, which may reflect the well-known reliance on external suppliers. It is reassuring, however, that its knowledge score remains high despite this.

Table 5 Characteristics by Knowledge Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Market Equity	6.55	6.36	6.25	6.08	5.95
Log Book Equity	5.93	5.66	5.35	5.21	5.07
Log Total Assets	6.8	6.44	6.06	5.82	5.65
Log Intangible Capital	6.27	6.1	6.02	5.81	5.62
Log Capital-Labor Ratio	4.75	4.75	4.67	4.72	4.9
Book-to-Market	0.727	0.691	0.591	0.58	0.605
Tobin's Q	1.38	1.8	2.32	2.49	2.68
Intangible-Adjusted Book-to-Market	2.49	2.39	2.15	2.03	1.9
Total Intangible Capital / Assets	1.11	1.23	1.5	1.48	1.19
Sales / Total Intangible Capital	7.35	2.67	1.42	1.39	1.38
Profitability	0.0736	0.0356	-0.0127	-0.0846	-0.171
Asset Tangibility	0.552	0.512	0.428	0.409	0.399
Solow Residual	0.0421	0.076	0.105	0.0117	-0.226
Intangible Capital Investment Rate	0.259	0.289	0.303	0.313	0.324
Physical Capital Investment Rate	0.102	0.138	0.158	0.164	0.171
Firm Age	28.4	18.3	15.6	16.5	17.7
Sales / Assets	1.18	1.04	0.926	0.813	0.648
Cash / Assets	0.12	0.18	0.262	0.32	0.403
R&D Expense / Assets	0.0147	0.0586	0.104	0.155	0.223
Advertising Expense / Assets	0.0305	0.0324	0.0319	0.0283	0.034
Debt / Assets	0.273	0.264	0.215	0.206	0.186
SG&A Expense / Assets	0.254	0.305	0.374	0.398	0.377
Employees / Assets	0.00724	0.00563	0.00472	0.00434	0.00322
SG&A Expense / Sales	0.24	0.436	0.722	1.11	2.29
R&D Expense / Sales	0.0232	0.206	0.461	2.72	12.3
Advertising Expense / Sales	0.0307	0.0335	0.0387	0.033	0.0429
Revenue per Employee	457	460	416	394	352
Labor Expense per Employee	78.1	73.9	76.8	102	130
Executive Compensation / Assets	2.71	3.66	4.7	5.07	3.98
θ_{int}	0.0854	0.143	0.21	0.251	0.3
θ_{know}	0.00499	0.0933	0.241	0.389	0.623
θ_{cust}	0.414	0.408	0.377	0.305	0.184
θ_{org}	0.501	0.467	0.365	0.3	0.191

Notes: This table presents the equal-weighted average of firm-level statistics for portfolios sorted on θ_{know}. All sorts are done within the Fama–French 17 industry classification. The sample period is from 2002 to 2023.

Table 6 Characteristics by Customer Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Market Equity	6	6.23	6.36	6.34	6.26
Log Book Equity	5.22	5.39	5.52	5.51	5.57
Log Total Assets	5.87	6.03	6.23	6.24	6.38
Log Intangible Capital	5.52	5.88	6.09	6.13	6.19
Log Capital-Labor Ratio	4.87	4.83	4.72	4.68	4.71
Book-to-Market	0.657	0.603	0.602	0.61	0.719
Tobin's Q	2.39	2.37	2.18	2.05	1.7
Intangible-Adjusted Book-to-Market	1.98	1.91	2.08	2.23	2.74
Total Intangible Capital / Assets	1.1	1.19	1.38	1.44	1.4
Sales / Total Intangible Capital	3.45	2.09	1.8	1.79	2.05
Profitability	-0.112	-0.0903	-0.022	0.0152	0.0474
Asset Tangibility	0.475	0.437	0.448	0.443	0.499
Solow Residual	-0.211	-0.118	0.0327	0.147	0.147
Intangible Capital Investment Rate	0.3	0.312	0.302	0.295	0.275
Physical Capital Investment Rate	0.151	0.163	0.145	0.146	0.127
Firm Age	19.8	18.2	19.3	18.7	20.4
Sales / Assets	0.769	0.783	0.909	0.991	1.15
Cash / Assets	0.307	0.324	0.253	0.233	0.171
R&D Expense / Assets	0.204	0.175	0.114	0.0885	0.0527
Advertising Expense / Assets	0.0237	0.0256	0.0288	0.0325	0.0397
Debt / Assets	0.21	0.214	0.233	0.227	0.258
SG&A Expense / Assets	0.319	0.344	0.352	0.358	0.328
Employees / Assets	0.00479	0.0045	0.00496	0.0048	0.00552
SG&A Expense / Sales	1.21	1.32	0.713	0.567	0.354
R&D Expense / Sales	9.07	6.71	1.38	0.24	0.0842
Advertising Expense / Sales	0.0261	0.0331	0.0347	0.0365	0.0395
Revenue per Employee	349	371	417	454	488
Labor Expense per Employee	111	103	81.2	64	61.8
Executive Compensation / Assets	3.6	4.02	4.22	4.09	3.6
θ_{int}	0.201	0.247	0.204	0.197	0.143
θ_{know}	0.355	0.387	0.29	0.228	0.0989
θ_{cust}	0.0402	0.174	0.306	0.45	0.713
θ_{org}	0.481	0.433	0.402	0.32	0.187

Notes: This table presents the equal-weighted average of firm-level statistics for portfolios sorted on θ_{cust}. All sorts are done within the Fama–French 17 industry classification. The sample period is from 2002 to 2023.

Table 7 Characteristics by Organization Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Market Equity	5.97	6.19	6.22	6.32	6.49
Log Book Equity	5.23	5.3	5.36	5.54	5.78
Log Total Assets	5.9	5.95	6.03	6.27	6.6
Log Intangible Capital	5.77	5.96	5.98	6.08	6.07
Log Capital-Labor Ratio	4.8	4.76	4.75	4.75	4.74
Book-to-Market	0.688	0.588	0.596	0.642	0.678
Tobin's Q	2.05	2.41	2.38	2.07	1.73
Intangible-Adjusted Book-to-Market	2.43	2.14	2.07	2.19	2.13
Total Intangible Capital / Assets	1.33	1.39	1.4	1.32	1.07
Sales / Total Intangible Capital	1.75	1.52	1.65	2.32	5.91
Profitability	-0.0387	-0.0816	-0.0558	-0.0176	0.0408
Asset Tangibility	0.458	0.408	0.434	0.478	0.524
Solow Residual	-0.0165	0.0115	0.00799	0.0465	-0.0072
Intangible Capital Investment Rate	0.29	0.312	0.303	0.293	0.283
Physical Capital Investment Rate	0.134	0.164	0.156	0.144	0.128
Firm Age	20.3	17.4	17.8	19.8	21.2
Sales / Assets	0.907	0.816	0.861	0.968	1.05
Cash / Assets	0.268	0.336	0.297	0.229	0.159
R&D Expense / Assets	0.142	0.167	0.145	0.11	0.0462
Advertising Expense / Assets	0.0354	0.0331	0.031	0.0276	0.0303
Debt / Assets	0.211	0.205	0.216	0.24	0.27
SG&A Expense / Assets	0.341	0.389	0.375	0.337	0.264
Employees / Assets	0.00461	0.00384	0.00425	0.00541	0.00667
SG&A Expense / Sales	0.625	0.858	0.915	0.702	0.396
R&D Expense / Sales	3.07	5.55	3.32	1.54	0.214
Advertising Expense / Sales	0.0365	0.0368	0.0382	0.0317	0.0298
Revenue per Employee	392	410	409	434	441
Labor Expense per Employee	88.1	103	87.7	85	81.4
Executive Compensation / Assets	3.71	4.56	4.27	3.98	3.28
θ_{int}	0.2	0.262	0.229	0.178	0.122
θ_{know}	0.383	0.403	0.312	0.197	0.0635
θ_{cust}	0.444	0.396	0.366	0.321	0.163
θ_{org}	0.0496	0.194	0.319	0.48	0.772

Notes: This table presents the equal-weighted average of firm-level statistics for portfolios sorted on θ_{org}. All sorts are done within the Fama–French 17 industry classification. The sample period is from 2002 to 2023.

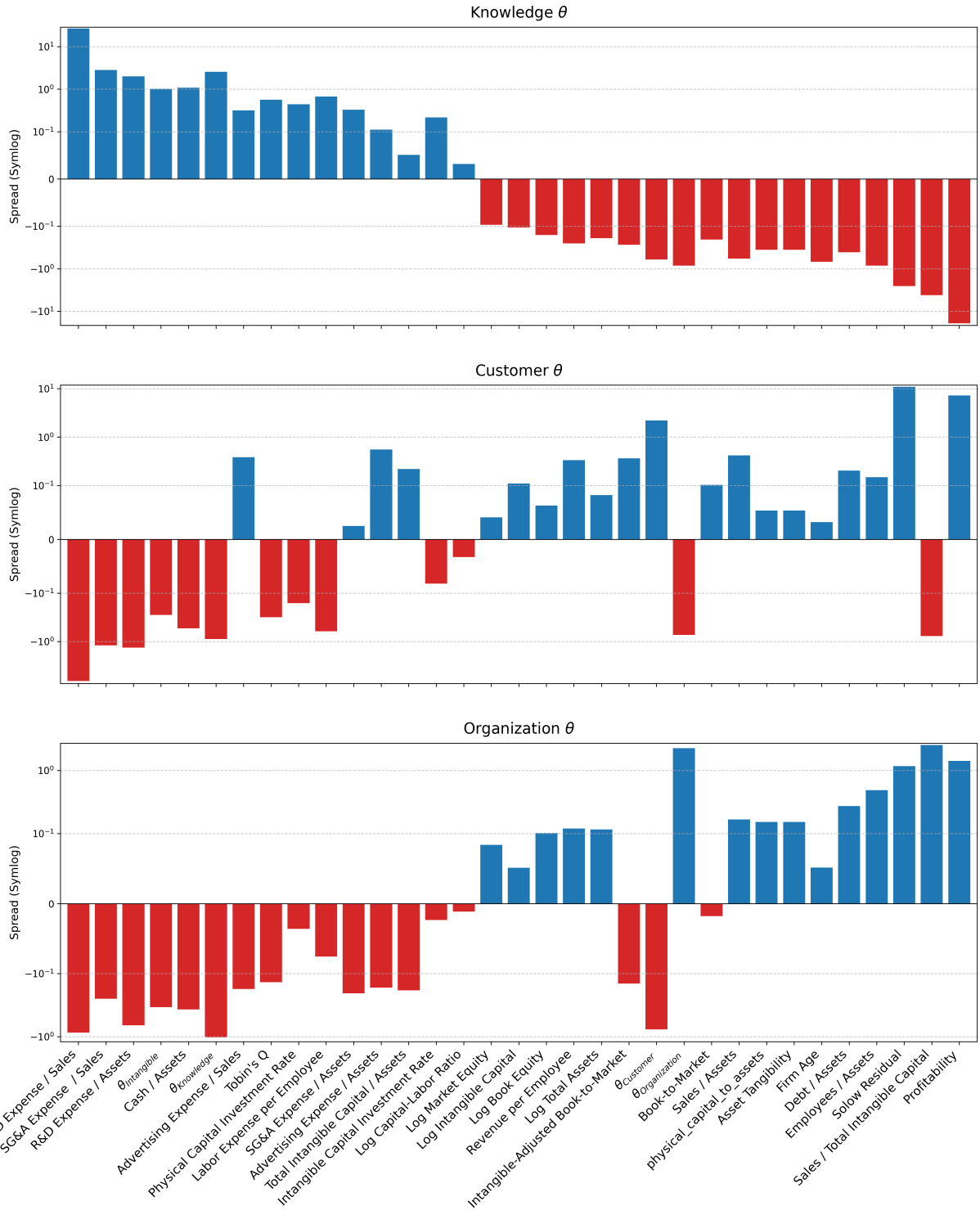


Figure 7: Q5-Q1 firm characteristics by θ

Notes: This figure reports the average difference of firm characteristics between Q5 and Q1 sorted firms by θ_{know}, θ_{cust} and θ_{org}.

5 Conclusion

This paper develops a new, text-based measure of intangible investment. We extract semantic information from firms' 10-K disclosures using a hybrid pipeline that combines LLM filtering and an embeddings model. The resulting measure, θ_{int}, captures the intensity with which firms articulate investment-related activities in their filings. We further decompose this firm-level intensity measure into knowledge, customer, and organization capital. The scores and their components exhibit substantial differences across industries and considerable heterogeneity within industries, and they are highly persistent over time.

Our empirical analysis shows that θ_{int} contains rich economic structure. Firms with high overall intangible intensity are smaller, younger, and in the midst of intensive investment characterized by high R&D and SG&A expenditures, elevated labor costs, and substantial liquidity needs. Decomposing θ_{int} reveals distinct economic profiles for each type of intangible capital: knowledge-intensive firms resemble classic R&D-driven growth firms; customer-intensive firms appear to be mature, profitable businesses focused on customer acquisition and market reach; and organization-intensive firms resemble operationally complex incumbents that rely on scale rather than frontier innovation. Relative to prior firm-level measures that assume fixed capitalization rates or rely on the subset of firms that voluntarily report specific expenditures, our approach provides independent variation that illuminates how these three forms of intangibles interact with one another and collectively trace out a firm's position in the lifecycle.

These measures open several avenues for future research. Because the scores are available at the firm-year level for nearly all public firms, they can be used to study whether intangible intensity predicts future operating performance, investment rates, or stock returns beyond what standard accounting ratios capture. The component-level decomposition also enables tests of how the composition of intangible investment relates to firm valuation, financing decisions, and exposure to aggregate risk. More broadly, θ_{int} and its subcomponents provide a scalable input for any setting in which researchers need a firm-level proxy for intangible investment that does not rely on the availability or accuracy of specific accounting line items.

Beyond the application to intangibles, our text-processing pipeline itself offers a methodological contribution. By combining LLM-based filtering with embedding-driven clustering, the framework can be adapted to other textual corpora and classification tasks, offering efficiency and accuracy gains over traditional dictionary or bag-of-words approaches while preserving interpretability.

References

Ahci, Mustafa and Philip Joos, "Text-Based Innovation Measure and Firm Performance," Technical Report 4797745, SSRN 2024.

Akcigit, Ufuk and Sina T. Ates, "What Happened to U.S. Business Dynamism?," Working Paper 25756, National Bureau of Economic Research 2019.

_ **and** _ , "Ten Facts on Declining Business Dynamism and Lessons from Endogenous Growth Theory," *American Economic Journal: Macroeconomics*, 2021, *13* (1), 257–298.

Alexander, Lewis and Janice C. Eberly, "Investment Hollowing Out," *IMF Economic Review*, 2018, *66* (1).

Amenc, Noël, Felix Goltz, and Ben Luyten, "Intangible Capital and the Value Factor: Has Your Value Definition Just Expired?," *The Journal of Portfolio Management*, 2020, *46* (7), 83–99.

Andrei, Daniel, William Mann, and Nathalie Moyen, "Why Did the Q Theory of Investment Start Working?," *Journal of Financial Economics*, 2019, *133* (2), 251–272.

Anton, James J. and Dennis A. Yao, "Little Patents and Big Secrets: Managing Intellectual Property," *RAND Journal of Economics*, 2004, *35* (1), 1–22.

Arnott, Robert D., Campbell R. Harvey, Vitali Kalesnik, and Juhani T. Linnainmaa, "Reports of Value's Death May Be Greatly Exaggerated," *Financial Analysts Journal*, 2021, *77* (1), 44–67.

Asness, Clifford S., R. Burt Porter, and Ross L. Stevens, "Predicting Stock Returns Using Industry-Relative Firm Characteristics," Technical Report 213872, SSRN 2000.

Atkeson, Andrew, "Alternative Facts Regarding the Labor Share," *Review of Economic Dynamics*, 2020, *37*, S167–S180.

_ **and Patrick J. Kehoe**, "Modeling and Measuring Organization Capital," *Journal of Political Economy*, 2005, *113* (5), 1026–1053.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 2016, *131* (4), 1593–1636.

Barkai, Simcha, "Declining Labor and Capital Shares," *The Journal of Finance*, 2020, *75* (5), 2421–2463.

Basu, Susanto, John G. Fernald, Nicholas Oulton, and Sylaja Srinivasan, "The Case of the Missing Productivity Growth," *NBER Macroeconomics Annual*, 2003, *18*, 9–63.

Belo, Frederico, Vito Gala, Juliana Salomao, and Maria Ana Vitorino, "Decomposing Firm Value," Working Paper, National Bureau of Economic Research 2019.

Benmelech, Efraim, "Asset Salability and Debt Maturity: Evidence from Nineteenth-Century American Railroads," *The Review of Financial Studies*, 2009, *22* (4), 1545–1584.

Bhide, A. V., *The Origin and Evolution of New Businesses*, Oxford University Press, 1999.

Bloom, Nicholas and John Van Reenen, "Measuring and Explaining Management Practices Across Firms and Countries," *The Quarterly Journal of Economics*, 2007, *122* (4), 1351–1408.

Bound, John, Clint Cummins, Zvi Griliches, Bronwyn H. Hall, and Adam B. Jaffe, "Who Does R&D and Who Patents?," Working Paper, National Bureau of Economic Research 1982.

Bresnahan, Timothy F., Erik Brynjolfsson, and Lorin M. Hitt, "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence," *The Quarterly Journal of Economics*, 2002, *117* (1), 339–376.

Bronnenberg, Bart J., Jean-Pierre Dubé, and Chad Syverson, "Marketing Investment and Intangible Brand Capital," *Journal of Economic Perspectives*, 2022, *36* (3), 53–74.

_ , Jean-Pierre H. Dubé, and Chad Syverson, "Intangible Marketing Capital," Working Paper 30145, National Bureau of Economic Research 2022.

Bybee, J. Leland, "The Ghost in the Machine: Generating Beliefs with Large Language Models," 2023. Working Paper.

Chan, Minki, Geonhee Hong, Joachim Hubmer, Serdar Ozkan, and Sergio Salgado, "Scalable versus Productive Technologies," Working Paper 2024-019, Federal Reserve Bank of St. Louis 2025.

Chandler, Alfred D., *The Visible Hand: The Managerial Revolution in American Business*, Harvard University Press, 1993.

Chen, Andrew J., Gerard Hoberg, and Miao Ben Zhang, "Haven't We Seen This Before? Return Predictions from 200 Years of News," Technical Report 5380343, SSRN 2025.

Cohen, Wesley M. and Daniel A. Levinthal, "Innovation and Learning: The Two Faces of R&D," *The Economic Journal*, 1989, *99* (397), 569–596.

Conrad, Jennifer, Michael Cooper, and Gautam Kaul, "Value versus Glamour," *The Journal of Finance*, 2003, *58* (5), 1969–1995.

Corrado, Carol, "Data and Intangible Asset Prices," Working Paper 2021.

Corrado, Carol A. and Charles R. Hulten, "How Do You Measure a "Technological Revolution"?," *American Economic Review*, 2010, *100* (2), 99–104.

Corrado, Carol, Charles Hulten, and Daniel Sichel, "Measuring Capital and Technology: An Expanded Framework," in "Measuring Capital in the New Economy," University of Chicago Press, 2005, pp. 11–46.

_ , _ , and _ , "Intangible Capital and U.S. Economic Growth," *Review of Income and Wealth*, 2009, *55* (3), 661–685.

_ , John Haltiwanger, and Daniel Sichel, eds, *Measuring Capital in the New Economy*, University of Chicago Press, 2009.

_ , Jonathan Haskel, Cecilia Jona-Lasinio, and Massimiliano Iommi, "Intangible Capital and Modern Economies," *Journal of Economic Perspectives*, 2022, *36* (3), 3–28.

_ , _ , Massimiliano Iommi, Cecilia Jona-Lasinio, and Filippo Bontadini, "Data, Intangible Capital, and Productivity," in "Technology, Productivity, and Economic Growth," University of Chicago Press, 2025.

Crouzet, Nicolas and Janice C. Eberly, "Intangibles, Investment, and Efficiency," *American Economic Review, Papers and Proceedings*, 2018, *108*, 426–431.

_ and _ , "Understanding Weak Capital Investment: The Role of Market Power and Intangibles," Jackson Hole Symposium Paper, Federal Reserve Bank of Kansas City 2019.

_ and _ , "Intangibles, Markups, and the Measurement of Productivity Growth," *Journal of Monetary Economics*, 2021. Forthcoming.

_ and _ , "Rents and Intangible Capital: A Q+ Framework," *The Journal of Finance*, 2023.

_ , _ , **Andrea L. Eisfeldt, and Dimitris Papanikolaou**, "The Economics of Intangible Capital," *Journal of Economic Perspectives*, 2022, *36* (3), 29–52.

Ding, Xiang, Teresa C. Fort, Stephen J. Redding, and Peter K. Schott, "Structural Change Within versus Across Firms: Evidence from the United States," Working Paper 2022.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart, "How to Make Causal Inferences Using Texts," *Science Advances*, 2022, *8* (42), eabg2652.

Eisfeldt, Andrea L. and Dimitris Papanikolaou, "Organization Capital and the Cross-Section of Expected Returns," *The Journal of Finance*, 2013, *68* (4), 1365–1406.

_ **and** _ , "The Value and Ownership of Intangible Capital," *American Economic Review*, 2014, *104* (5), 189–194.

_ **and Gregor Schubert**, "Generative AI and Finance," *Annual Review of Financial Economics*, 2025, *17* (1), 363–393.

_ **, Antonio Falato, and Mindy Z. Xiaolan**, "Human Capitalists," Working Paper 28815, National Bureau of Economic Research 2021.

_ **, Edward T. Kim, and Dimitris Papanikolaou**, "Intangible Value," *Critical Finance Review*, 2022, *11* (2), 299–332.

Elsby, Michael W. L., Bart Hobijn, and Ayşegül Şahin, "The Decline of the U.S. Labor Share," *Brookings Papers on Economic Activity*, 2013, *2013* (2), 1–63.

Enache, Luminita and Anup Srivastava, "Should Intangible Investments Be Reported Separately or Commingled with Operating Expenses? New Evidence," *Management Science*, 2018, *64* (7), 3446–3468.

Ewens, Michael, Ryan H. Peters, and Sean Wang, "Measuring Intangible Capital with Market Prices," *Management Science*, 2024, *71* (1), 407–427.

Falato, Antonio, Dalida Kadyrzhanova, Jae Sim, and Roberto Steri, "Rising Intangible Capital, Shrinking Debt Capacity, and the U.S. Corporate Savings Glut," *The Journal of Finance*, 2022, *77* (5), 2799–2852.

Faria, André L., "Mergers and the Market for Organization Capital," *Journal of Economic Theory*, 2008, *138* (1), 71–100.

Gentzkow, Matthew and Jesse M. Shapiro, "What Drives Media Slant? Evidence from U.S. Daily Newspapers," *Econometrica*, 2010, *78* (1), 35–71.

_ , Bryan Kelly, and Matt Taddy, "Text as Data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Gomme, Paul, B. Ravikumar, and Peter Rupert, "The Return to Capital and the Business Cycle," *Review of Economic Dynamics*, 2011, *14* (2), 262–278.

Goncalves, Andrei and Gregory Leonard, "The Fundamental-to-Market Ratio and the Value Premium Decline," Working Paper, Kenan Institute of Private Enterprise 2020.

Gorton, Gary B., Jillian Grennan, and Alexander K. Zentefis, "Corporate Culture," *Annual Review of Financial Economics*, 2022, *14* (1), 535–561.

Gourio, François and Leena Rudanko, "Customer Capital," *Review of Economic Studies*, 2014, *81* (3), 1102–1136.

Graham, John R., Jillian Grennan, Campbell R. Harvey, and Shivaram Rajgopal, "Corporate Culture: Evidence from the Field," *Journal of Financial Economics*, 2022, *146* (2), 552–593.

Gutiérrez, Germán and Thomas Philippon, "Investment-Less Growth: An Empirical Investigation," Working Paper, National Bureau of Economic Research 2016.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg, "Market Value and Patent Citations," *RAND Journal of Economics*, 2005, *36* (1), 16–38.

Hansen, Lars Peter, John C. Heaton, and Nan Li, "Intangible Risk," in "Measuring Capital in the New Economy," University of Chicago Press, 2009, pp. 111–152.

Hart, Oliver and Bengt Holmstrom, "A Theory of Firm Scope," *The Quarterly Journal of Economics*, 2010, *125* (2), 483–513.

_ and John Moore, "A Theory of Debt Based on the Inalienability of Human Capital," *The Quarterly Journal of Economics*, 1994, *109* (4), 841–879.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, "Firm-Level Political Risk: Measurement and Effects," *The Quarterly Journal of Economics*, 2019, *134* (4), 2135–2202.

He, Huan (Bianca), Lauren Mostrom, and Amir Sufi, "Investing in Customer Capital," Working Paper 2024-144, University of Chicago, Becker Friedman Institute for Economics 2024.

Helmers, Christian and Mark Rogers, "Does Patenting Help High-Tech Start-Ups?," *Research Policy*, 2011, *40* (7), 1016–1027.

Hoberg, Gerard and Gordon M. Phillips, "Scope, Scale and Concentration: The 21st Century Firm," *Journal of Finance*, 2024. Forthcoming.

_ **and Gordon Phillips**, "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 2016, *124* (5), 1423–1465.

Hulten, Charles and Xiaohui Hao, "What is a Company Really Worth? Intangible Capital and the Market to Book Value Puzzle," Working Paper 14548, National Bureau of Economic Research 2008.

Ide, Enrique and Eduard Talamas, "Artificial Intelligence in the Knowledge Economy," *Journal of Political Economy*, 2025. Forthcoming.

Iqbal, Aneel, Shiva Rajgopal, Anup Srivastava, and Rong Zhao, "A Better Estimate of Internally Generated Intangible Capital," *Management Science*, 2025, *71* (1), 731–752.

Jones, Charles I., "R&D-Based Models of Economic Growth," *Journal of Political Economy*, 1995, *103* (4).

_ **and Christopher Tonetti**, "Nonrivalry and the Economics of Data," *American Economic Review*, 2020, pp. 2819–2858.

Jovanovic, Boyan and Peter L. Rousseau, "The Q-Theory of Mergers," *American Economic Review*, 2002, *92* (2), 198–204.

Karabarbounis, Loukas and Brent Neiman, "The Global Decline of the Labor Share," *The Quarterly Journal of Economics*, 2014, *129* (1), 61–103.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, "Measuring Technological Innovation Over the Long Run," *American Economic Review: Insights*, 2021, *3* (3), 303–320.

Kermani, Amir and Yueran Ma, "Asset Specificity of Non-Financial Firms," Working Paper 27642, National Bureau of Economic Research 2020.

Kogan, Leonid and Dimitris Papanikolaou, "Growth Opportunities, Technology Shocks, and Asset Prices," *The Journal of Finance*, 2014, *69* (2), 675–718.

_ , _ , Lawrence D. W. Schmidt, and Bryan Seegmiller, "Technology-Skill Complementarity and Labor Displacement: Evidence from Linking Two Centuries of Patents with Occupations," Working Paper 29552, National Bureau of Economic Research 2021.

Lashkari, Danial, Arthur Bauer, and Jocelyn Boussard, "Information Technology and Returns to Scale," Technical Report 3458604, SSRN 2018.

Lev, Baruch, *Intangibles: Management, Measurement, and Reporting*, Brookings Institution Press, 2000.

_ and Anup Srivastava, "Explaining the Recent Failure of Value Investing," Working Paper, NYU Stern School of Business 2019.

_ and Suresh Radhakrishnan, "The Valuation of Organization Capital," in "Measuring Capital in the New Economy," University of Chicago Press, 2005, pp. 73–110.

Loecker, Jan De, Jan Eeckhout, and Gabriel Unger, "The Rise of Market Power and the Macroeconomic Implications," *The Quarterly Journal of Economics*, 2020, *135* (2), 561–644.

Lucas, Robert E. and Benjamin Moll, "Knowledge Growth and the Allocation of Time," *Journal of Political Economy*, 2014, *122* (1), 1–51.

Lustig, Hanno, Chad Syverson, and Stijn Van Nieuwerburgh, "Technological Change and the Growing Inequality in Managerial Compensation," *Journal of Financial Economics*, 2011, *99* (3), 601–627.

Markovitch, Dmitri G., Dongling Huang, and Pengfei Ye, "Marketing Intensity and Firm Performance: Contrasting the Insights Based on Actual Marketing Expenditure and Its SG&A Proxy," *Journal of Business Research*, 2020, *118*, 223–239.

Michaelides, Alexander, Andreas Milidonis, Vitaliy Ryabinin, and Yupana Wiwattanakantang, "The Value of Trade Secrets: Evidence from Economic Espionage," Technical Report 4866808, SSRN 2024.

Ottonello, Pablo, Wenting Song, and Sebastian Sotelo, "An Anatomy of Firms' Political Speech," Working Paper w32923, National Bureau of Economic Research 2024.

Papanikolaou, Dimitris, "Investment Shocks and Asset Prices," *Journal of Political Economy*, 2011, *119* (4), 639–685.

Park, Hyuna, "An Intangible-Adjusted Book-to-Market Ratio Still Predicts Stock Returns," *Critical Finance Review*, 2022, *11* (2), 265–297.

Peters, Ryan H. and Lucian A. Taylor, "Intangible Capital and the Investment-q Relation," *Journal of Financial Economics*, 2017, *123* (2), 251–272.

Philippon, Thomas, "The Bond Market's q," *The Quarterly Journal of Economics*, 2009, *124* (3), 1011–1056.

Rajgopal, Shivaram, Anup Srivastava, and Rong Zhao, "Do Digital Technology Firms Earn Excess Profits? Alternative Perspectives," *The Accounting Review*, 2023, *98* (4), 321–344.

Ramey, Valerie A. and Matthew D. Shapiro, "Displaced Capital: A Study of Aerospace Plant Closings," *Journal of Political Economy*, 2001, *109* (5), 958–992.

Ridder, Maarten De, "Market Power and Innovation in the Intangible Economy," Working Paper 2022.

Rizova, Savina and Namiko Saito, "Intangibles and Expected Stock Returns," Technical Report 3697452, SSRN 2020.

Smith, Matthew, Danny Yagan, Owen Zidar, and Eric Zwick, "Capitalists in the Twenty-First Century," *The Quarterly Journal of Economics*, 2019, *134* (4), 1675–1745.

Solow, Robert M., "The Production Function and the Theory of Capital," *The Review of Economic Studies*, 1955, pp. 103–107.

_ , "Technical Change and the Aggregate Production Function," *The Review of Economics and Statistics*, 1957, *39* (3), 312–320.

Srivastava, Anup, "Why Have Measures of Earnings Quality Changed Over Time?," *Journal of Accounting and Economics*, 2014, *57* (2), 196–217.

_ , "Trivialization of the Bottom Line and Losing Relevance of Losses," *Review of Accounting Studies*, 2023, *28*, 1190–1208.

Sun, Qi and Mindy Z. Xiaolan, "Financing Intangible Capital," *Journal of Financial Economics*, 2019, *133* (3), 564–588.

Syverson, Chad, "Macroeconomics and Market Power: Context, Implications, and Open Questions," *Journal of Economic Perspectives*, 2019, *33* (3), 23–43.

Tambe, Prasanna, Lorin Hitt, Daniel Rock, and Erik Brynjolfsson, "Digital Capital and Superstar Firms," Working Paper 28285, National Bureau of Economic Research 2020.

Wang, Xizhao, "Patenting and Information Disclosure," Technical Report 5387619, SSRN 2025.

Weiss, Joshua, "Intangible Investment and Market Concentration," Working Paper, New York University 2019.

Zarifhonarvar, Ali, "Generating Inflation Expectations with Large Language Models," *Journal of Monetary Economics*, 2025, p. 103859.

Zhang, Mindy Xiaolan, "Who Bears Firm-Level Risk? Implications for Cash Flow Volatility," 2013. Working Paper.

Appendix A Additional Details on Methodology

This appendix provides additional details on the methodology for extracting 10-K text and classifying n-grams into relevant communities.

To extract the relevant quotes from raw MD&A text, we use the below prompt:

```
QUOTE_EXTRACTION_PROMPT_TEMPLATE = """
You are a precise, rule-following financial analyst. Your task is to analyze text
    from a company's 10-K and extract the most relevant sentences that explain a
    specific expense category.

EXPENSE CATEGORY TO ANALYZE:
{component_name}

CONTEXT TO ANALYZE:
The following text snippets have been pre-selected because they likely mention "{
    component_name}" and discuss financial results.
---
{context_blocks}
---

CRITICAL INSTRUCTIONS (MUST BE FOLLOWED):

1.  Extract Three Types of Information: Your goal is to extract direct quotes that
     fall into three specific categories.

        1. Definition Quotes (The "What"):
         -  Extract sentences that define the composition of the expense. These
             sentences describe what types of costs make up the expense category (e.
             g., personnel, marketing, facilities).

        2. Business Driver Quotes (The "Ongoing Why"):
         -   Extract sentences that explain the fundamental business purpose of the
              expense. These sentences describe the core activities the expense
             supports and why the company incurs these costs as a part of its
             strategy.

        3. Change Driver Quotes (The "Why it Changed"):
           -  Extract sentences that explain the specific reasons for an increase or
                decrease in the expense during the reporting period. These sentences
               often compare the current period to a prior period and mention
               specific causes for the variance.

2.  Extract Direct Quotes: You MUST extract the information as direct quotes from
    the "CONTEXT TO ANALYZE". Each quote MUST be a complete, full sentence from its
     beginning to its end (e.g., to the period). Do not extract partial phrases or
    fragments. Do not paraphrase.

3.  No Data Fabrication (Most Important Rule): If you cannot find any relevant
    quotes for a specific category in the provided context, you MUST return an
    empty list for that category (e.g., `"change_driver_quotes": []`). Do not
    invent information or use any text from these instructions.

OUTPUT SCHEMA:
Your entire response MUST be a single, valid JSON object that follows the
    structure below. Do not add any text before or after the JSON object.

{{
  "component_name": "{component_name}",
  "definition_quotes": [
  ],
  "business_driver_quotes": [
  ],
  "change_driver_quotes": [
  ]
}}
"""
"""
```

In constructing the n-gram universe, we begin with all bigrams and trigrams extracted from the *LLM-processed texts*, restricted to noun–noun and noun–noun–noun patterns. The full master list includes tens of thousands of n-grams, but for tractability we focus on the top $N = 10{,}000$ by frequency; expanding the set to $N = 20{,}000$ reduces the frequency of the least common n-gram from 7 to 4 and produces similar results. Each n-gram is embedded using the OpenAI

text-embedding-3-large model,[14] producing an E-dimensional vector.

Next, we apply a sequence of transformations: (i) centering the embedding matrix by subtracting the mean vector, (ii) projecting onto the first P principal components, (iii) whitening the projected vectors using the inverse square root of the eigenvalue matrix, and (iv) L2-normalizing each row. These steps correct the anisotropy often present in transformer embedding spaces and reduce the influence of noisy dimensions. Clustering proceeds by applying K-Means to the normalized embedding vectors, producing K micro-clusters with centroids c_k. To construct communities, we build a k-nearest-neighbor similarity graph on the centroids, where edges are drawn when cosine similarity exceeds a threshold δ. The Leiden algorithm partitions this graph into 247 communities, which are then hand-labeled. Formally, each n-gram n_j inherits the label of the community containing the micro-cluster C_k for which $n_j \in C_k$. The Online Appendix also reports diagnostics such as the proportion of variance explained by PCA, sensitivity of community counts to alternative K, and comparisons of clustering stability across different random seeds.

[14] Our approach does not rely on using a large embedding model; results are robust to smaller open-source models.

Appendix B Additional Figures and Tables

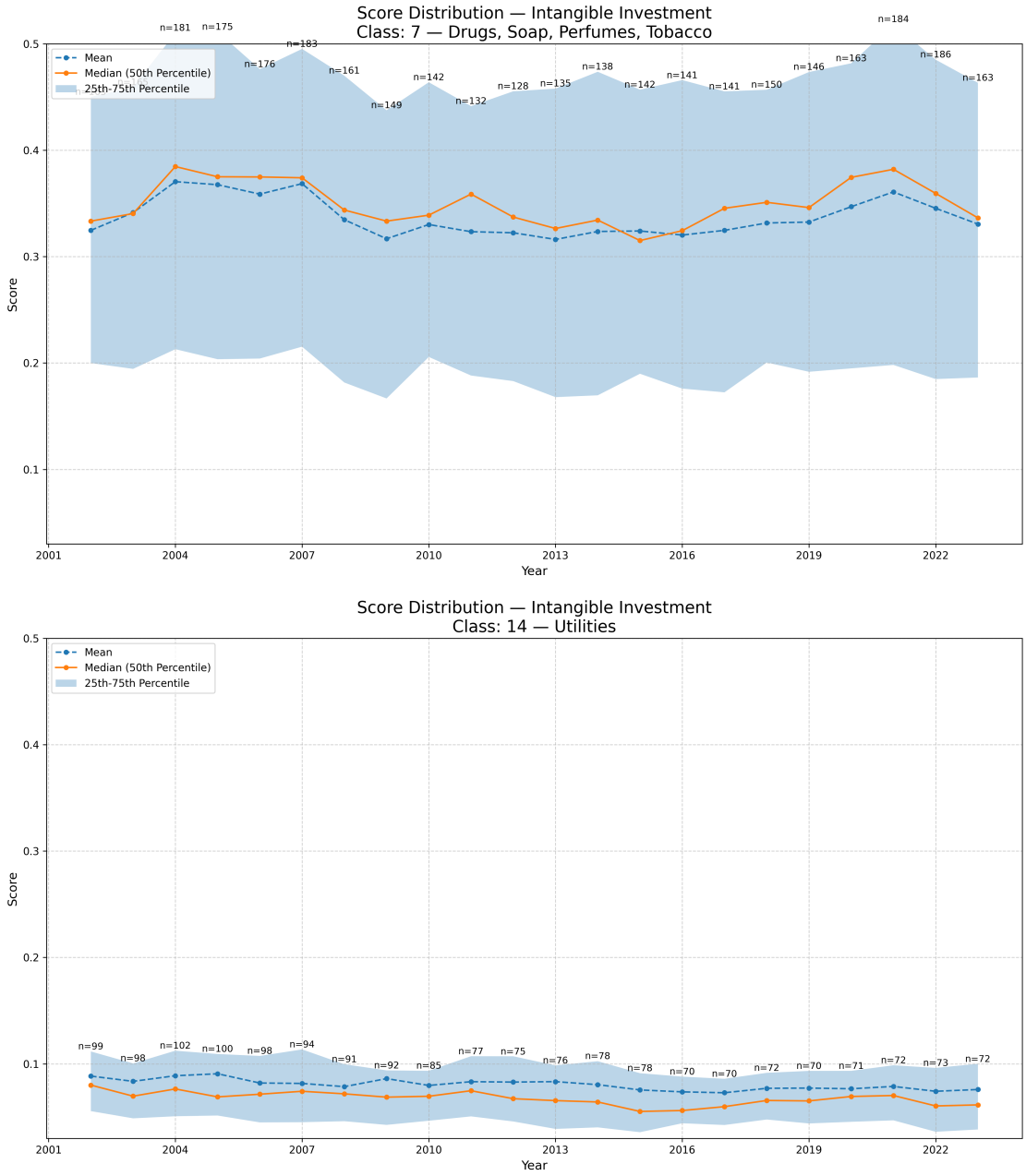


Figure B.1: Intangible Intensity Distribution for Select Industries

Notes: This figure reports the average θ_{int} and its subcomponents for the "Drugs, Soaps, Perfumes, and Tobacco" industry and the "Utilities" industry.

Table B.1 Example Community Assignments and Categories

community_id	size	mean_similarity	representatives	category_name	subcategory_name
5	168	0.6265	compensation cost expense, compensation expense cost, service compensation expense, compensation expense amount, time compensation expense, compensation expense number	Unknown	
6	167	0.5176	expense office rent, cost office rent, rent expense office, office rental expense, expense rent office	Not intangible investment	
4	145	0.5601	advertising promotion expense, advertising expense promotion, advertising expense cost, advertising promotion cost, advertising cost expense, show advertising expense	Intangible investment	Customer capital
0	142	0.7118	cost commission sale, commission sale expense, commission level sale, commission expense cost, sale commission expense	Intangible investment	Customer capital
2	140	0.6490	percent sale percentage, expense percent percentage, product percentage sale, sale percent percentage, percentage cost sale, cost percentage sale	Unknown	
44	140	0.7002	function research development, term research development, research development work, value research development, research development project, research development solution	Intangible investment	Knowledge capital

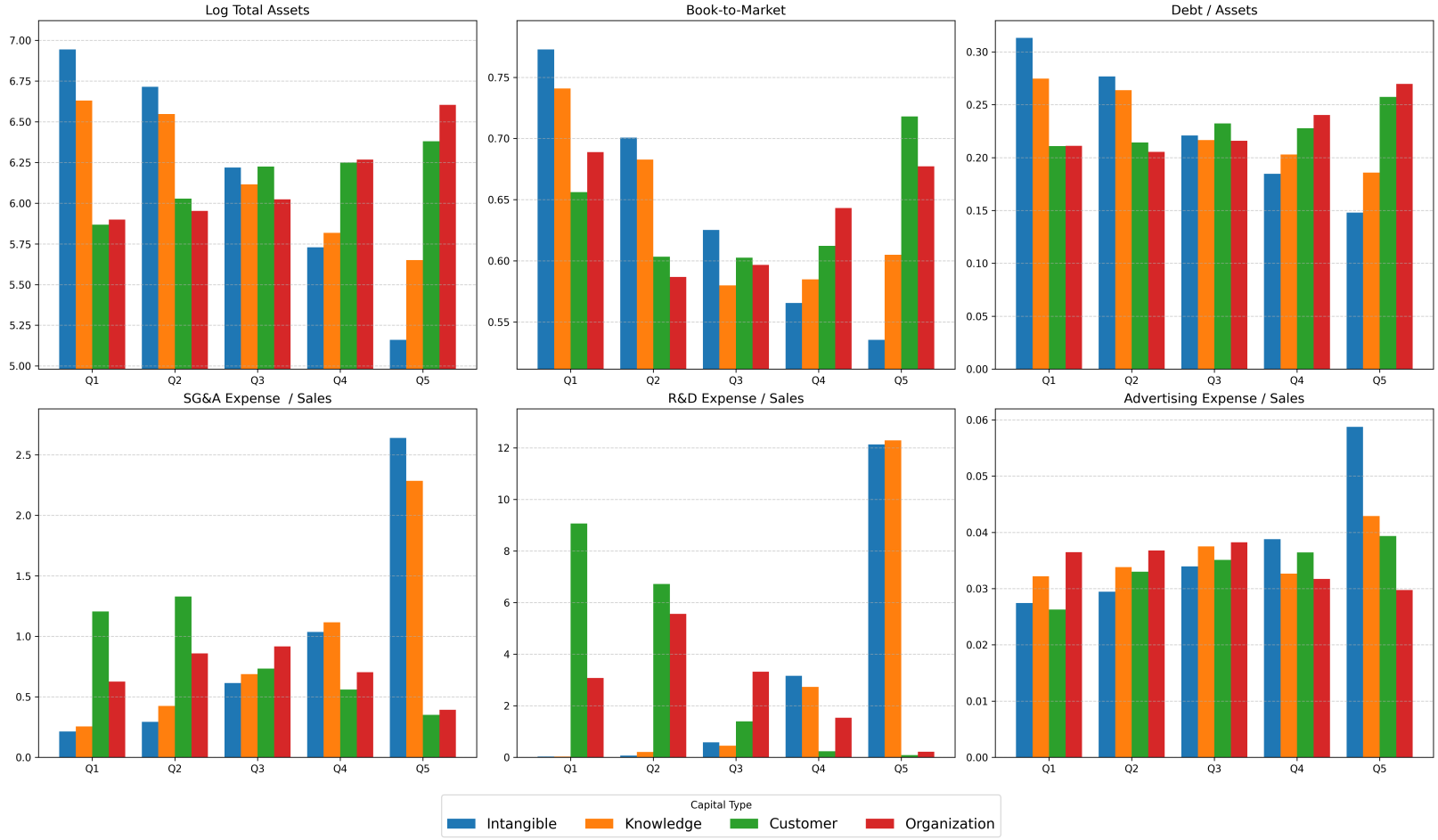


Figure B.2: Time-Averaged Values Across Quintiles

Notes: This figure reports the time-averaged selected variables across quintiles over different θ sorts.

Table B.2 Industry Average Intensity Summary Statistics (FF17)

Industry	Type	Mean	SD	P25	P50	P75
Food	$\theta_{\text{Intangible}}$	0.125	0.082	0.070	0.110	0.163
Food	θ_{Customer}	0.484	0.289	0.250	0.500	0.714
Food	$\theta_{\text{Knowledge}}$	0.079	0.154	0.000	0.000	0.100
Food	$\theta_{\text{Organization}}$	0.406	0.283	0.182	0.385	0.633
Mining and Minerals	$\theta_{\text{Intangible}}$	0.090	0.093	0.033	0.063	0.114
Mining and Minerals	θ_{Customer}	0.332	0.308	0.044	0.250	0.500
Mining and Minerals	$\theta_{\text{Knowledge}}$	0.076	0.173	0.000	0.000	0.064
Mining and Minerals	$\theta_{\text{Organization}}$	0.535	0.332	0.297	0.538	0.840
Oil and Petroleum Products	$\theta_{\text{Intangible}}$	0.060	0.049	0.028	0.050	0.079
Oil and Petroleum Products	θ_{Customer}	0.298	0.294	0.000	0.236	0.500
Oil and Petroleum Products	$\theta_{\text{Knowledge}}$	0.060	0.151	0.000	0.000	0.036
Oil and Petroleum Products	$\theta_{\text{Organization}}$	0.573	0.337	0.333	0.636	0.857
Textiles, Apparel & Footwear	$\theta_{\text{Intangible}}$	0.110	0.072	0.060	0.094	0.146
Textiles, Apparel & Footwear	θ_{Customer}	0.508	0.265	0.328	0.500	0.687
Textiles, Apparel & Footwear	$\theta_{\text{Knowledge}}$	0.066	0.120	0.000	0.000	0.092
Textiles, Apparel & Footwear	$\theta_{\text{Organization}}$	0.402	0.269	0.200	0.400	0.600
Consumer Durables	$\theta_{\text{Intangible}}$	0.143	0.107	0.065	0.111	0.205
Consumer Durables	θ_{Customer}	0.434	0.282	0.222	0.400	0.600
Consumer Durables	$\theta_{\text{Knowledge}}$	0.201	0.238	0.000	0.105	0.372
Consumer Durables	$\theta_{\text{Organization}}$	0.322	0.266	0.091	0.300	0.500
Chemicals	$\theta_{\text{Intangible}}$	0.131	0.120	0.053	0.092	0.166
Chemicals	θ_{Customer}	0.322	0.248	0.131	0.286	0.463
Chemicals	$\theta_{\text{Knowledge}}$	0.273	0.257	0.000	0.250	0.456
Chemicals	$\theta_{\text{Organization}}$	0.371	0.262	0.167	0.333	0.548
Drugs, Soap, Perfumes, Tobacco	$\theta_{\text{Intangible}}$	0.339	0.173	0.193	0.351	0.474
Drugs, Soap, Perfumes, Tobacco	θ_{Customer}	0.264	0.227	0.094	0.200	0.362
Drugs, Soap, Perfumes, Tobacco	$\theta_{\text{Knowledge}}$	0.500	0.276	0.304	0.566	0.721
Drugs, Soap, Perfumes, Tobacco	$\theta_{\text{Organization}}$	0.230	0.179	0.106	0.189	0.304
Construction and Construction Materials	$\theta_{\text{Intangible}}$	0.100	0.075	0.049	0.084	0.133
Construction and Construction Materials	θ_{Customer}	0.417	0.303	0.167	0.400	0.641
Construction and Construction Materials	$\theta_{\text{Knowledge}}$	0.085	0.163	0.000	0.000	0.105
Construction and Construction Materials	$\theta_{\text{Organization}}$	0.452	0.303	0.200	0.467	0.667
Steel Works Etc	$\theta_{\text{Intangible}}$	0.076	0.057	0.042	0.069	0.104
Steel Works Etc	θ_{Customer}	0.350	0.292	0.075	0.333	0.556
Steel Works Etc	$\theta_{\text{Knowledge}}$	0.103	0.168	0.000	0.000	0.167
Steel Works Etc	$\theta_{\text{Organization}}$	0.423	0.315	0.167	0.400	0.636
Fabricated Products	$\theta_{\text{Intangible}}$	0.097	0.065	0.050	0.078	0.127
Fabricated Products	θ_{Customer}	0.424	0.287	0.190	0.394	0.633
Fabricated Products	$\theta_{\text{Knowledge}}$	0.156	0.219	0.000	0.008	0.250
Fabricated Products	$\theta_{\text{Organization}}$	0.391	0.292	0.167	0.333	0.602

Continued on next page

Continued from previous page

Industry	Type	Mean	SD	P25	P50	P75
Machinery and Business Equipment	$\theta_{\text{Intangible}}$	0.208	0.130	0.102	0.188	0.295
Machinery and Business Equipment	θ_{Customer}	0.320	0.217	0.167	0.289	0.438
Machinery and Business Equipment	$\theta_{\text{Knowledge}}$	0.378	0.250	0.171	0.400	0.559
Machinery and Business Equipment	$\theta_{\text{Organization}}$	0.288	0.230	0.118	0.238	0.410
Automobiles	$\theta_{\text{Intangible}}$	0.132	0.103	0.061	0.105	0.170
Automobiles	θ_{Customer}	0.354	0.257	0.155	0.315	0.500
Automobiles	$\theta_{\text{Knowledge}}$	0.278	0.269	0.000	0.222	0.468
Automobiles	$\theta_{\text{Organization}}$	0.346	0.260	0.143	0.300	0.519
Transportation	$\theta_{\text{Intangible}}$	0.105	0.078	0.051	0.086	0.142
Transportation	θ_{Customer}	0.339	0.290	0.091	0.286	0.500
Transportation	$\theta_{\text{Knowledge}}$	0.125	0.216	0.000	0.000	0.167
Transportation	$\theta_{\text{Organization}}$	0.499	0.331	0.229	0.500	0.800
Utilities	$\theta_{\text{Intangible}}$	0.081	0.064	0.044	0.068	0.101
Utilities	θ_{Customer}	0.325	0.254	0.118	0.302	0.500
Utilities	$\theta_{\text{Knowledge}}$	0.125	0.193	0.000	0.033	0.175
Utilities	$\theta_{\text{Organization}}$	0.502	0.287	0.295	0.500	0.714
Retail Stores	$\theta_{\text{Intangible}}$	0.105	0.072	0.054	0.091	0.137
Retail Stores	θ_{Customer}	0.417	0.278	0.200	0.400	0.611
Retail Stores	$\theta_{\text{Knowledge}}$	0.032	0.088	0.000	0.000	0.000
Retail Stores	$\theta_{\text{Organization}}$	0.519	0.293	0.296	0.536	0.750
Other	$\theta_{\text{Intangible}}$	0.242	0.168	0.096	0.216	0.363
Other	θ_{Customer}	0.327	0.241	0.130	0.302	0.480
Other	$\theta_{\text{Knowledge}}$	0.315	0.282	0.000	0.286	0.538
Other	$\theta_{\text{Organization}}$	0.335	0.268	0.135	0.255	0.500

Table B.3 Variable Definitions and Formulas

Variable	**Definition**
Log Market Equity	$\log(\text{me})$
Log Book Equity	$\log(\text{be})$
Log Total Assets	$\log(\text{at})$
Log Intangible Capital	$\log(\text{k}^*_{int})$
Log Capital-Labor Ratio	$\log(\frac{ppegt}{emp})$
Book-to-Market	$\frac{\text{be}}{\text{me}}$
Tobin's Q	$\frac{\text{me+dltt+dlc+cshpri}-\text{invt}}{\text{at}}$
ROA	$\frac{\text{ib}}{\text{at}_{-1}}$
Profitability	$\frac{\text{ib+dp}}{\text{at}_{-1}}$
Asset Tangibility	$\frac{\text{ppegt}}{\text{at}}$
Solow Residual	solow residual*
Firm Age	age*
Intangible Capital Investment Rate	$\frac{\text{xsga}}{\text{k}_{int,-1}}$
Physical Capital Investment Rate	$\frac{\text{capx}}{\text{ppegt}_{-1}}$
Total Intangible Capital / Assets	$\frac{\text{k}_{int}}{\text{at}}$
Intangible-Adjusted Book-to-Market	$\frac{\text{be+k}_{int}-\text{gdwl}}{me}$
Sales / Total Intangible Capital	$\frac{\text{sale}}{\text{k}_{int,-1}}$
Sales / Assets	$\frac{\text{sale}}{\text{at}}$
Cash / Assets	$\frac{\text{che}}{\text{at}}$
R&D Expense / Assets	$\frac{\text{xrd}}{\text{at}}$
Advertising Expense / Assets	$\frac{\text{xad}}{\text{at}}$
Debt / Assets	$\frac{\text{dltt+dlc}}{\text{at}}$
SG&A Expense / Assets	$\frac{\text{xsga}}{\text{at}}$
Employees / Assets	$\frac{\text{emp}}{\text{at}}$
Executive Compensation / Assets	$\frac{\text{tdc2}}{\text{at}}$
SG&A Expense / Sales	$\frac{\text{xsga}}{\text{sale}}$
R&D Expense / Sales	$\frac{\text{xrd}}{\text{sale}}$
Advertising Expense / Sales	$\frac{\text{xad}}{\text{sale}}$
Revenue per Employee	$\frac{\text{sale}}{\text{emp}}$
Labor Expense per Employee	$\frac{\text{xlr}}{\text{emp}}$

Notes: k_{int} is calculated using perpetual inventory method and capitalizes 100% of xsga using a depreciation rate of 20%. age is calculated as the corresponding year in panel minus the firm's birth year, where we approximate birth year as the minimum of first appearance in CRSP, first appearance in Compustat, and ipodate. solow residual is computed as the residual of a regression of log sales on log physical capital and log labor at industry-year level. x_{-1} indicates a lagged variable.

Table B.4 Intangibles Intensity for Selected Firms

Company	Ticker	SG&A/Sales	Median	R&D/Sales	Med	Adv/Sales	Median	θ_{int}	θ_{know}	θ_{cust}	θ_{org}
Apple Inc.	AAPL	12.0%	47.4%	6.0%	20.8%		1.9%	1	4	3	3
Microsoft	MSFT	27.3%	47.4%	12.3%	20.8%	0.9%	1.9%	4	3	5	1
Alphabet (Google)	GOOG	26.4%	47.4%	12.3%	20.8%	3.1%	1.9%	2	2	5	3
Amazon	AMZN	36.7%	24.4%	11.9%	0.0%	3.6%	2.2%	4	4	4	2
Tesla	TSLA	13.2%	14.5%	4.8%	3.7%		0.9%	4	4	3	2
Nvidia	NVDA	27.6%	26.3%	19.6%	7.7%		0.6%	3	5	4	1
Meta Platforms	META	41.1%	47.4%	20.9%	20.8%	2.5%	1.9%	3	4	4	2
Procter & Gamble	PG	27.4%	52.7%	2.5%	25.7%	10.8%	4.1%	1	1	3	5
Walmart	WMT	20.2%	24.4%	0.0%	0.0%	0.7%	2.2%	1	3	1	1
Eli Lilly & Co.	LLY	47.5%	52.7%	27.9%	25.7%	4.2%	4.1%	2	5	2	1
Pfizer Inc.	PFE	29.3%	52.7%	17.0%	25.7%	2.5%	4.1%	2	3	4	1
Home Depot	HD	16.6%	13.8%	0.0%	0.8%	0.7%	0.5%	1	2	1	5
Coca-Cola	KO	31.0%	21.1%		0.4%	10.6%	2.8%	4	4	2	4
AbbVie	ABBV	33.8%	47.4%	14.3%	20.8%	3.7%	1.9%	2	4	2	2
Merck & Co.	MRK	43.8%	52.7%	25.1%	25.7%	4.1%	4.1%	2	2	4	4
PepsiCo	PEP	39.1%	21.1%	0.9%	0.4%	4.4%	2.8%	4	5	3	3
Verizon	VZ	22.0%	47.4%		20.8%	2.5%	1.9%	2	2	5	4
Thermo Fisher	TMO	20.7%	26.3%	3.6%	7.7%		0.6%	1	3	5	1

Notes: Selected firms for FY 2021. θs refer to the firm's within-industry quintile assignment from 1 to 5. Median represents the industry median value for the preceding column.

Constructing θ_{int}: A Text-Based Measure of Intangible Intensity

Andrea L. Eisfeldt

Barney Hartman-Glaser

Edward T. Kim

Ki Beom Lee

NBER Digital Economics and AI Tutorial

February 2026

Motivation: The Problem

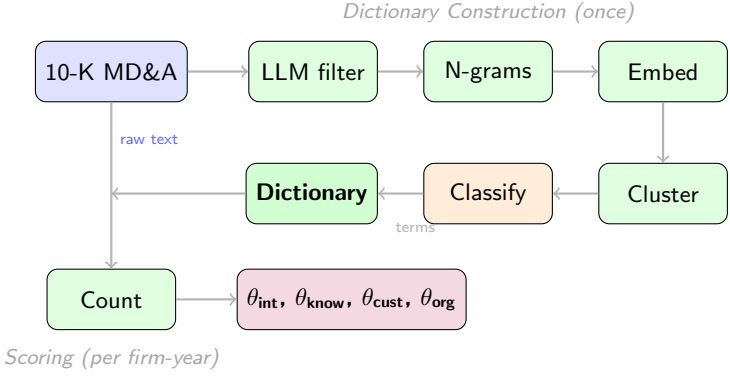
Intangible investment is large ($\sim 1/2$ or more of corporate capital)

But accounting measures are problematic:

- R&D reported by only $\sim$65% of firms
- Advertising reported by only $\sim$41% of firms
- No accounting proxy for organization capital
- SG&A mixes investment with operating costs

Our goal: Text-based measure that is scalable, interpretable, and broadly available

Pipeline: Embedding-Based Dictionary + Frequency Scoring



Front-loads semantic sophistication into dictionary construction and keeps scoring simple, fast, and auditable.

Step 1: Text Extraction

Source: Item 7 (MD&A) of Form 10-K filings

Problem: MD&A contains boilerplate (liquidity, accounting policies)

Solution: Use lightweight LLM (gemma3n) to extract SG&A-relevant passages

Prompt Sketch:

- **Role:** Financial analyst extracting quotes from 10-K MD&A sections
- **Input:** Pre-filtered text snippets around expense keywords (SG&A, G&A, Sales & Marketing, Operating Expenses)
- **Extracts 3 quote types:**
  1. **Definition** – what costs make up the expense
  2. **Business Drivers** – why the company incurs it
  3. **Change Drivers** – why it increased/decreased vs. prior period

Step 1: Text Extraction — Example

Raw MD&A excerpt (Amazon 2022):

> *"The increase in sales and marketing costs in absolute dollars in 2022, compared to the prior year, is primarily due to increased payroll and related expenses for personnel engaged in marketing and selling activities and higher marketing spend."*

What the LLM extracts:

- Business driver: "marketing and selling activities"
- Change driver: "increased payroll," "higher marketing spend"

Why this matters:

- Filters out boilerplate (liquidity discussion, accounting policy)
- Focuses dictionary construction on economically relevant text
- **LLM used for curation, NOT classification at this step**

Step 2: N-gram Dictionary

Process:

- Extract bigrams and trigrams with noun-only structure
- Frequency-rank to create "master list"
- Keep top $N = 10{,}000$ N-grams

Why N-grams?

- Transparent and interpretable
- Easy to audit and classify
- **Atomic semantic units (vs. averaging over paragraphs)**

Step 3: Embeddings & Post-Processing

Embed N-grams using sentence-transformer model (text-embedding-3-large)

Problem: Raw embeddings blur economically distinct phrases

- e.g., "software development services" $\approx$ "software license maintenance"

Post-processing pipeline:

1. **PCA:** Dimensionality reduction from 3072 to 100, remove noise,
2. **Whitening:** Normalize the variance of the PCs (corrects anisotropy)
3. **Normalization:** Unit length vectors for cosine similarity

Step 3: Why Post-Processing Matters

The anisotropy problem:

- Pre-trained embeddings cluster in a narrow cone
- Similarity is artificially high for unrelated phrases
- K-means fails because everything looks similar

Step	Problem it solves
PCA	Removes noise dimensions, keeps signal
Whitening	Spreads vectors evenly in all directions
Normalization	Makes cosine similarity meaningful

Practical tip: Don't skip this. Raw embeddings $\rightarrow$ garbage clusters.

Step 4: Clustering & Community Detection

The scale problem:

- 10,000 N-grams is too many to hand-label individually
- But simple K-means with reasonable k gives arbitrary boundaries

Two-stage approach:

1. **K-Means:** Create fine-grained micro-clusters from embeddings
2. **Leiden algorithm:** Merge micro-clusters into coherent communities

Result: 247 communities of semantically related N-grams

Validation: Relative to random groupings in a Monte Carlo benchmark, our communities show higher within-community cosine similarity and smaller average distance to the community centroid.

Step 5: Classification

Classify each community into 5 categories:

- **Knowledge capital** (R&D, technology development)
- **Brand or customer capital** (marketing, brand, sales)
- **Organization capital** (operations, processes, management)
- **Non-intangible** (routine expenses, facilities)
- **Unknown**

Intangible = Knowledge $\cup$ Customer $\cup$ Organization

Note: Can also classify *communities* with an LLM

Step 5: Example Communities

The classification task:

- Look at the N-grams in each community
- Ask: Does this represent intangible investment activity?
- If yes: Which type (knowledge, customer, organization)?

ID	Size	Sample N-grams	Class
44	140	"research development work", "research development project"	Knowledge
4	145	"advertising promotion expense", "advertising cost expense"	Customer
0	142	"cost commission sale", "commission sale expense"	Customer
6	167	"expense office rent", "cost office rent"	Non-intang.
5	168	"compensation cost expense", "service compensation expense"	Unknown

Step 6: Scoring Documents

For each firm-year:

1. Extract N-grams from raw *Item 7: Management Discussion and Analysis (MD&A)* text (not LLM-processed)
2. Restrict to whitelist W (the 10,000 classified N-grams)
3. Count N-grams by category

$$\theta_{\text{int}} = \frac{\text{intangible N-gram count}}{\text{total whitelist N-gram count}}$$

Similarly for subcategories: θ_{know}, θ_{cust}, θ_{org}

Interpretation: Share of relevant text devoted to *intangible investment* (see Hassan et al. 2025 on text-based economic measurement)

Step 6: Scoring — The Details

Why score on raw text, not LLM-processed?

- LLM-processed text is concentrated SG&A discussion
- Would artificially inflate intangible proportions
- Raw text = fair representation of what firm emphasizes overall

Why restrict to whitelist?

- Apples-to-apples comparison across firms
- Unclassified N-grams (not in top 10K) are excluded
- Ensures consistent denominator

Interpretation of θ

Semantic attention: Share of MD&A devoted to intangible-related activities

Investment intensity: Proxy for fraction of SG&A that represents intangible investment

Summary statistics:

Measure	Mean	SD	P25	P50	P75
θ_{int}	0.194	0.156	0.071	0.143	0.291
θ_{know}	0.263	0.277	0.000	0.182	0.480
θ_{cust}	0.342	0.257	0.143	0.308	0.500
θ_{org}	0.367	0.283	0.143	0.300	0.560

Summary of Methodology

Step	Method	Output
1. Text extraction	LLM (gemma3n)	Filtered SG&A text
2. N-gram corpus	Frequency filtering	10,000 N-grams
3. Embeddings	Sentence-transformer + PCA/whitening	Dense vectors
4. Clustering	K-Means + Leiden	247 communities
5. Dictionary classification	Hand- or model-labeling	5 categories
6. Scoring	N-gram counting	θ_{int}, θ_{know}, θ_{cust}, θ_{org}

Key Innovation:

- Relative to alternatives, our method expands semantic sophistication into dictionary construction but keeps scoring simple, fast, and auditable
- Avoids treating LLMs as per-document oracles while substantially improving on traditional bag-of-words approaches

Three Types of Intangible Firms

The subcategories capture fundamentally different firm profiles:

Characteristic	High θ_{know}	High θ_{cust}	High θ_{org}
Firm type	Growth/R&D	Mature/Commercial	Operational scale
Profitability	Negative	Positive	Positive
Tobin's Q	High	Low	Low
Advertising/Sales	High	High	Low
R&D/Sales	High	Low	Low
Cash holdings	High	Low	Low
Leverage	Low	High	High

Correlation structure: Firms specialize in one intangible strategy

	θ_{know}	θ_{cust}	θ_{org}
θ_{know}	100%	-35%	-51%
θ_{cust}		100%	-44%

Why This Approach?

Takeaway: Semantic sophistication belongs upstream so that measurement can stay simple, transparent, and reproducible

Alternatives considered:

Approach	Method	Problem
Traditional dictionary	Hand-pick keywords	Can't scale to all 10K terms
GPT per-document	LLM classifies each 10-K	Not reproducible, expensive, garbage
Semantic axis	Embed poles, project docs	Noisy and intangibility isn't one axis
Soft matching	Seeds + embedding similarity	Threshold-dependent

Sample Code Available



https://github.com/bhglaser/semantic-theme-scoring