Intangible Intensity

Andrea L. Eisfeldt*

Barney Hartman-Glaser†

Edward T. Kim‡

Ki Beom Lee§

March 2026

LINK TO GITHUB CODE REPOSITORY

Abstract

We develop a text-based measure of intangible investment intensity derived from firms' 10-K filings, and offer a general methodology for semantic theme scoring (STS). 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 expenditure descriptions thus provide informative signals about intangible investment, complementing financial statements in capturing corporate capital stocks.

For helpful comments, the authors thank seminar participants at the NBER Digital Economics and AI Tutorial. *First draft: November 18, 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 more than half of corporate capital (Corrado et al., 2009; Eisfeldt and Papanikolaou, 2013, 2014; Peters and Taylor, 2017; Belo et al., 2022; 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 a co-product of the organization. In this paper, we develop a novel text-based measure of intangible investment intensity derived from firms' 10-K filings. Using a robust, 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. public firms. Our measure captures meaningful heterogeneity in firms' disclosures about intangible expenditures and provides an integrated framework to study how intangible investment shapes firm outcomes.

While accounting data contains information about intangible investment, precise measurement is notoriously difficult. Selling, general, and administrative (SG&A) expenses contain intangible spending but confound genuine investment with other expenses. In addition, research and development (R&D) and advertising are voluntary and underreported as separate items, while no line item exists for spending on organization capital, which is reported by Corrado et al. (2009) to be the largest category of corporate intangibles. Even when data are available, it is often unclear what fraction of expenditures reflects investment rather than recurring costs. A natural complement to the quantitative spending data are 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 black box language models, have prevented the systematic use of disclosures to measure intangible investment at scale.

We overcome these challenges by introducing *semantic theme scoring (STS)*, 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 ("Management Discussion and Analysis," or "MD&A") 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 30,000 high-frequency n-grams grouped into roughly 700 coherent communities, which are then classified into intangible and non-intangible categories. For the communities classified as intangible, we further decompose them into knowledge, brand or customer (henceforth customer), and organization 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. Our method covers nearly the entire universe of public firms from 2002 to 2023 and produces stable, persistent measures with substantial variation across industries, as well as considerable heterogeneity within industries.

We document important, intuitive variation in firm-level intangibles using θ^{int} and its components. Sorting firms each year into intangible-intensity 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 spending, 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 for a full universe of firms, unaffected by the sizeable and selective variation in firms' decisions about whether to report disaggregated intangible spending such as R&D, advertising, and selling and marketing expenses. 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 R&D spending, and heavy reliance on skilled labor. Customer capital-intensive firms, in contrast, exhibit characteristics of mature, profitable firms focused on customer acquisition and retention, and commercial scale, rather than innovation. Organization capital-intensive firms reflect a different dimension altogether: they are large, asset-intensive incumbents with substantial operational complexity. This component has been especially difficult to measure in prior work due to the absence of a targeted accounting proxy. 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 answers to 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 (1957) 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 (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). Supplementing Compustat with Capital IQ data improves customer capital expense measures (He et al., 2024), but there remains no systematic way to decompose SG&A consistently across all firms. 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 (Gentzkow and Shapiro, 2010; Baker et al., 2016; Hoberg and Phillips, 2016, 2024; Gentzkow et al., 2019; Ahci and Joos, 2024; Ottonello et al., 2024; Bybee, 2025; Clayton et al., 2025). Most early text analyses relied on bag-of-words or keyword searches using predefined vocabularies, including applications that measure the share of discussion devoted to particular topics (Hassan et al., 2019; Gentzkow et al., 2019). While robust, the reliance of these methods on researchers' dictionary inputs limits the ability to capture relevant text. On the other extreme, a fast-growing body of work in finance now uses LLMs directly to score or classify text (Clayton et al., 2025). These approaches can extract richer information than

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

dictionary methods, but the scoring process can be opaque, difficult to reproduce, and hard to scale with researcher expertise (Eisfeldt and Schubert, 2025). We address these concerns by using a hybrid approach that retains the transparency of dictionary-based methods while relying on LLMs for corpus selection and n-gram embeddings for measurement. Our LLM assists in constructing the dictionary, but scoring is deterministic and fully auditable.[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 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 useful but 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 (Lev and Radhakrishnan, 2005; Eisfeldt and Papanikolaou, 2014; Eisfeldt et al., 2022).[3] At the same time, SG&A also reflects other routine and non-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 short-run operational needs. 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 certain 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, while investments towards organization capital are not reported as a

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

[3]Lev and Radhakrishnan (2005) and Lev (2000) 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.

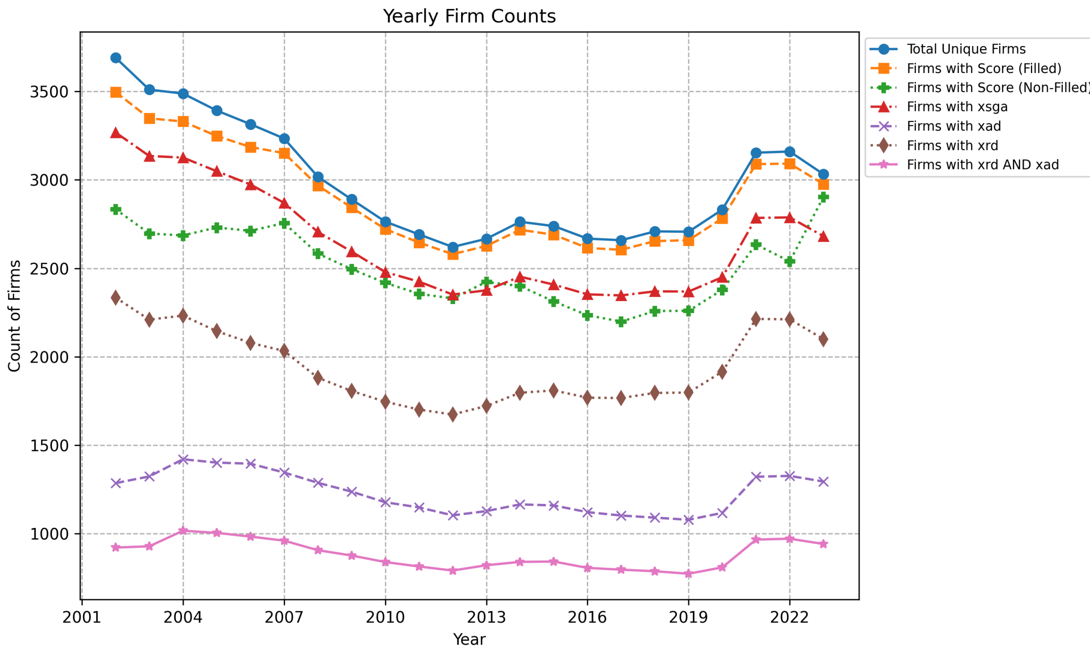


Figure 1: Coverage of Spending Variables

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

standalone item at all. 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. Furthermore, the omissions are selected, not idiosyncratic or random; 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.

2.2 Data Preparation

Our measure of intangible investment is constructed from Form 10-K filings, which provide narrative disclosures of firms' operating activities and expenses.[4] The sample includes all publicly traded firms from 2002 to 2023, the period over which digitized 10-K filings are readily available

[4]The STS process is described in more detail in Appendix A.1. A code repository replicating the pipeline is available at `https://github.com/bhglaser/semantic-theme-scoring`.

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 definitions are provided in the Appendix and associated Github repository.

We identify the portions of 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 10-K filings include an Item 7 MD&A section as standard practice, and most MD&A disclosures contain discussions of SG&A expenditures. As such, the MD&A provides a targeted chunk of text to analyze.

MD&A also contains discussions of liquidity management 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, or "whitelist", 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 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. 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.

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

[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.

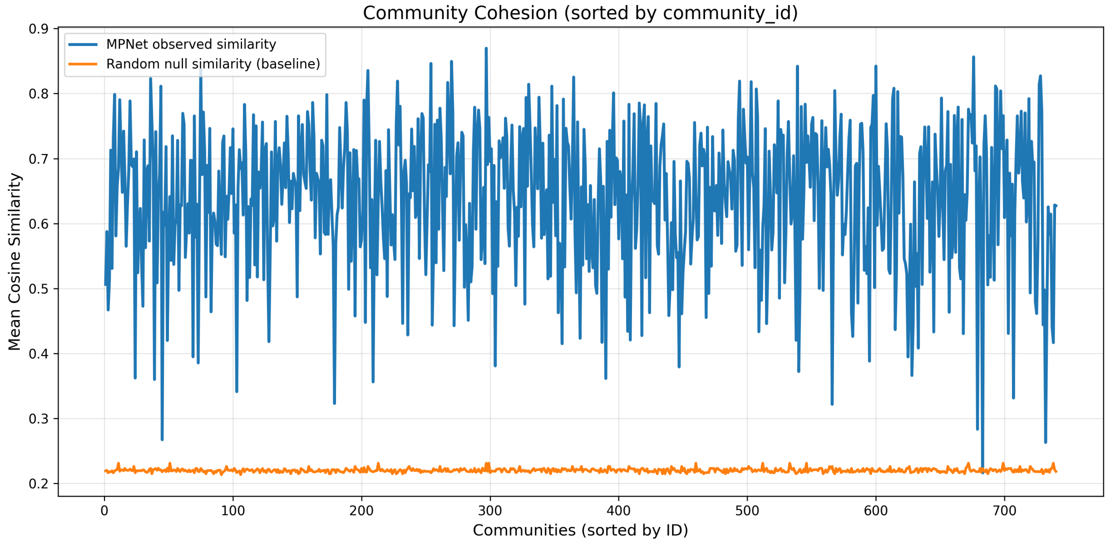


Figure 2: Community Cohesion Benchmarking

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

semantically related n-grams. This process yields 247 communities. Each community is then hand-classified into one of four categories: knowledge, customer, organization, and non-intangible. Combining algorithmic community identification and clustering with researcher classification effectively leverages the scalability enabled by modern LLMs with the domain knowledge of the researcher. While results are similar if this last step utilizes a large-scale LLM, the small number of communities will still allow the researcher to effectively monitor the process and output. The knowledge, customer, and organization categories collectively define the set of intangible-related n-grams.[6] These communities form the backbone of our dictionary. In summary, 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 embedding 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] An example set of communities and our classifications is reproduced in Appendix Table B.1.

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 two main categories (intangible, non-intangible), and if it is intangible, to one of the three subcategories (knowledge, customer, organization).

To construct scores, we extract all 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.[7] 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 discussed in MD&A. At the same time, restricting attention to the whitelist of n-grams ensures comparability across firms and avoids contamination from boilerplate or idiosyncratic language. This approach provides a stable basis for measurement.

Figure 3 illustrates the distribution of n-gram coverage in the raw texts. Because the whitelist W retains only noun-based, plain-English n-grams, the covered fraction is modest in absolute terms but represents non-trivial coverage of economically meaningful content. Coverage rises smoothly in the size of the n-gram universe, though at a declining rate as additional entries are drawn from increasingly infrequent terms. We adopt $N = 10{,}000$ as our baseline (robust to larger samples), which offers a good balance between coverage and tractability.

[7] 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.

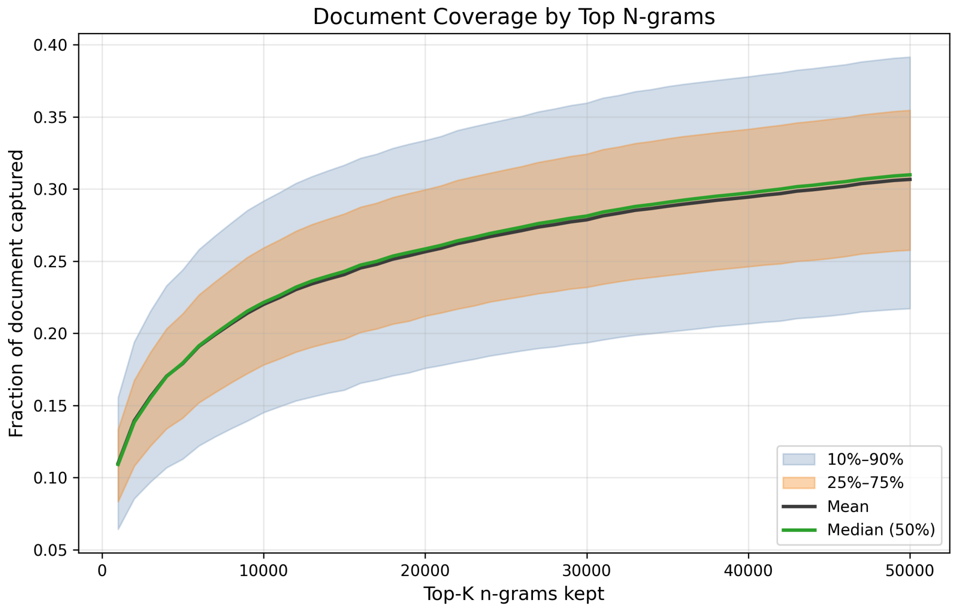


Figure 3: Whitelist Length and N-grams Coverage

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.

1. **Data:** For each document d, let $S(d)$ denote the set of all bigrams and trigrams extracted from the raw MD&A text, and define $A(d) := S(d) \cap W$ as the subset appearing in the whitelist. For each n-gram $n \in A(d)$, let $f_d(n) \in \mathbb{N}$ denote its frequency in the document.

2. **Aggregation:** For each category $c \in \mathcal{C} = \{\text{intangible, non-intangible}\}$, define the total frequency

$$F_d(c) = \sum_{n \in A(d):\, \text{cat}(n)=c} f_d(n), \tag{1}$$

and analogously for each subcategory $s \in \mathcal{S} = \{\text{know, cust, org}\}$,

$$F_d(s) = \sum_{n \in A(d):\, \text{sub}(n)=s} f_d(n). \tag{2}$$

By construction, $F_d(\text{intangible}) = \sum_{s \in \mathcal{S}} F_d(s)$. We can also define community-level aggregates $F_d(h) = \sum_{n \in A(d):\, \mathrm{comm}(n)=h} f_d(n)$, which are useful for industry-level analyses.

3. **Scores:** For each document d, the intangible intensity score is

$$\theta^{\text{int}} \;=\; \frac{F_d(\text{intangible})}{\sum_{c \in \mathcal{C}} F_d(c)}, \tag{3}$$

and the subcategory shares are defined over intangible n-grams only:

$$\theta^{\text{know}} \;=\; \frac{F_d(\text{know})}{\sum_{s \in \mathcal{S}} F_d(s)}, \quad \theta^{\text{cust}} \;=\; \frac{F_d(\text{cust})}{\sum_{s \in \mathcal{S}} F_d(s)}, \quad \theta^{\text{org}} \;=\; \frac{F_d(\text{org})}{\sum_{s \in \mathcal{S}} F_d(s)}. \tag{4}$$

By definition, $\theta^{\text{int}} + \theta^{\text{non-int}} + \theta^{\text{unk}} = 1$ and $\theta^{\text{know}} + \theta^{\text{cust}} + \theta^{\text{org}} = 1$.

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 intangible investment activities or expenses, consistent with evidence that textual emphasis tracks actual firm behavior (Hassan et al., 2019). 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.

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.

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.[8] This procedure preserves the time-series structure of the filing language and raises coverage to one

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

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.

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.

3.2 Industry Dynamics of Intensity

The wide distribution of intangible 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

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

exhibit lower values.

We document these facts in Figure 5, by plotting intensity measures across the Fama-French 17 industries.[9] 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).[10] In contrast, industries with 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 economic priors about across-industry differences in technological and organizational demands.

We also study how each intangible capital 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

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

[10] Firms in this industry likely exhibit substantial within-industry dispersion in terms of business segment and innovation strategies.

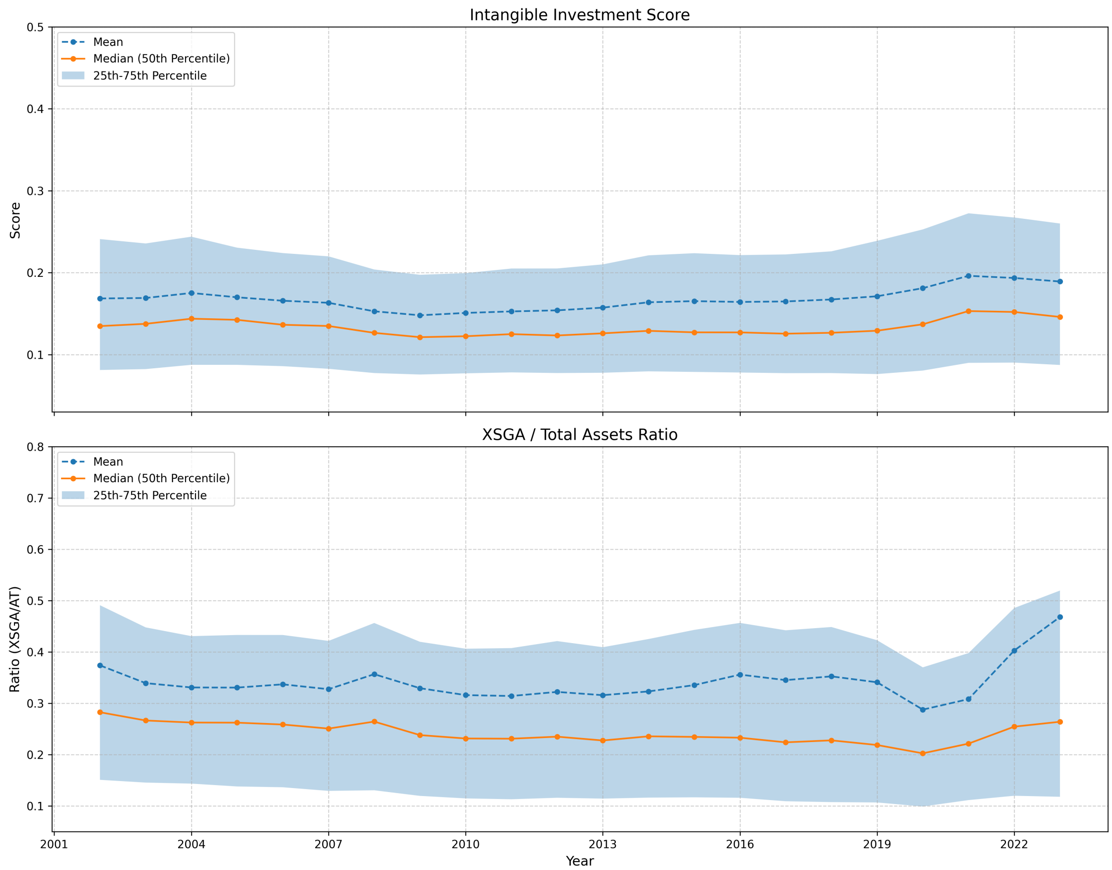


Figure 4: Distribution of Intangible Intensity Over Time

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

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 **??** 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.

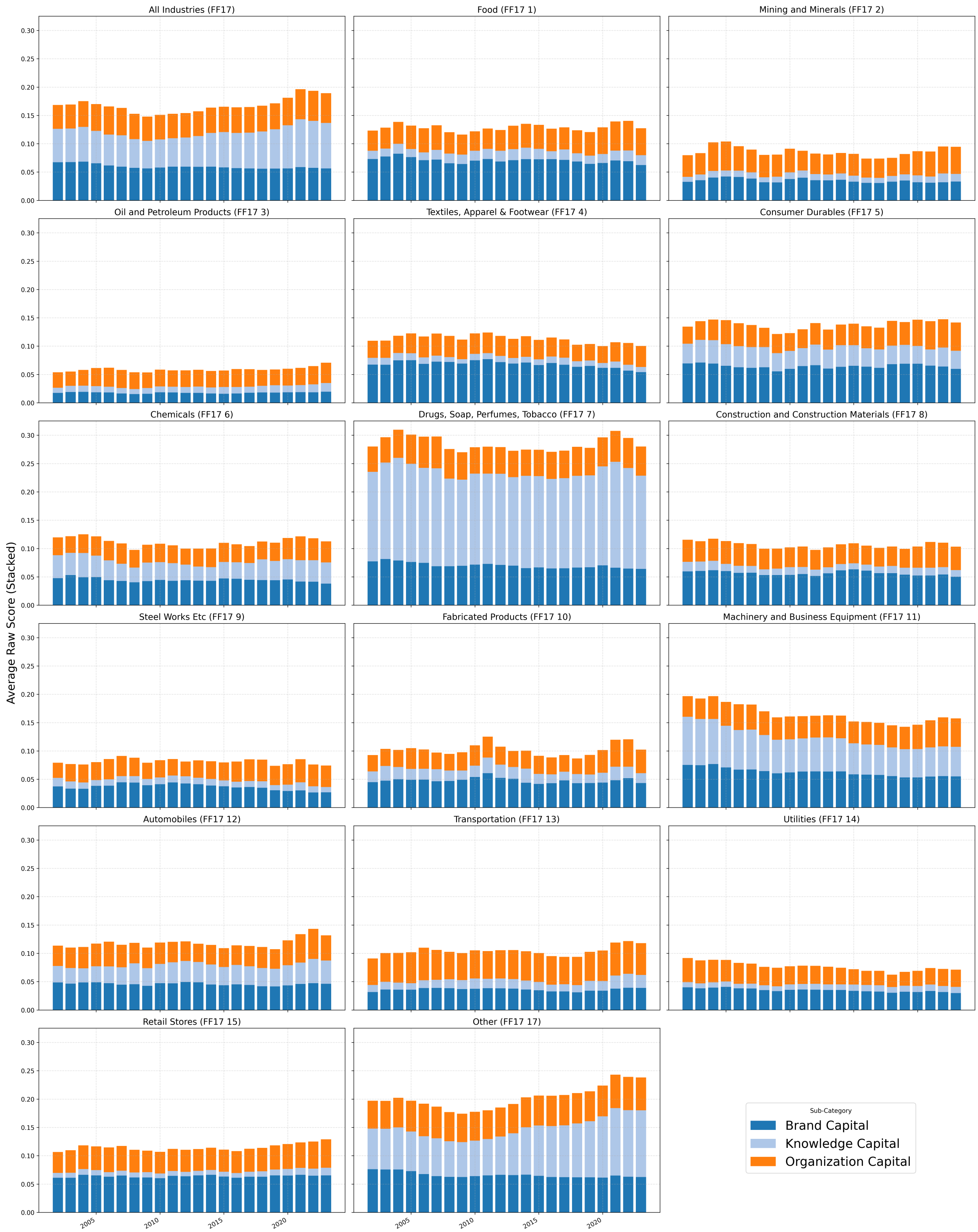


Figure 5: Intangible Intensity and Subcomponents Across Industries

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 **??** 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. High

Table 3 Characteristics by Intangible Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Market Equity	6.71	6.48	6.22	5.96	5.71
Log Book Equity	6.09	5.75	5.38	4.98	4.64
Log Total Assets	7.01	6.59	6.04	5.57	5.14
Log Intangible Capital	3.97	3.93	3.65	3.38	3.02
Intangible Capital Share	0.315	0.382	0.465	0.516	0.549
Knowledge Capital Share	0.334	0.334	0.333	0.329	0.313
Brand Capital Share	0.378	0.379	0.379	0.38	0.388
Book-to-Market	0.795	0.703	0.614	0.575	0.53
Tobin's Q	1.59	1.82	2.22	2.54	2.9
Profitability	0.119	0.101	0.0354	-0.049	-0.174
Asset Tangibility	0.285	0.25	0.208	0.181	0.148
Physical Capital Investment Rate	0.216	0.264	0.327	0.395	0.521
Sales / Assets	1.04	1.09	0.983	0.875	0.716
Cash / Assets	0.107	0.153	0.241	0.338	0.485
Debt / Assets	0.315	0.273	0.216	0.175	0.14
Employees / Assets	0.00619	0.00644	0.00538	0.00438	0.00316
Revenue per Employee	427	424	409	394	334
Labor Expense per Employee	58.7	85.1	76.9	92	128
Executive Compensation / Assets	2.59	3.35	4.33	5.56	6.78
R&D Expense / Assets	0.0183	0.0478	0.097	0.159	0.248
Advertising Expense / Assets	0.0254	0.0275	0.0286	0.0327	0.0458
SG&A Expense / Assets	0.203	0.279	0.36	0.43	0.468
SG&A Expense / Sales	0.216	0.324	0.733	1.1	1.97
R&D Expense / Sales	0.027	0.105	0.851	2.76	11.5
Advertising Expense / Sales	0.0261	0.0303	0.0335	0.0392	0.054
R&D Expense / SG&A Expense	0.029	0.0754	0.142	0.194	0.241
Advertising Expense / SG&A Expense	0.074	0.106	0.165	0.192	0.158
θ^{int}	0.0616	0.109	0.159	0.219	0.326
θ^{know}	0.139	0.198	0.267	0.334	0.447
θ^{cust}	0.397	0.41	0.406	0.385	0.325
θ^{org}	0.463	0.39	0.325	0.28	0.227

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. All firm characteristic variables are winsorized at the 1% level. The sample period is from 2002 to 2023.

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

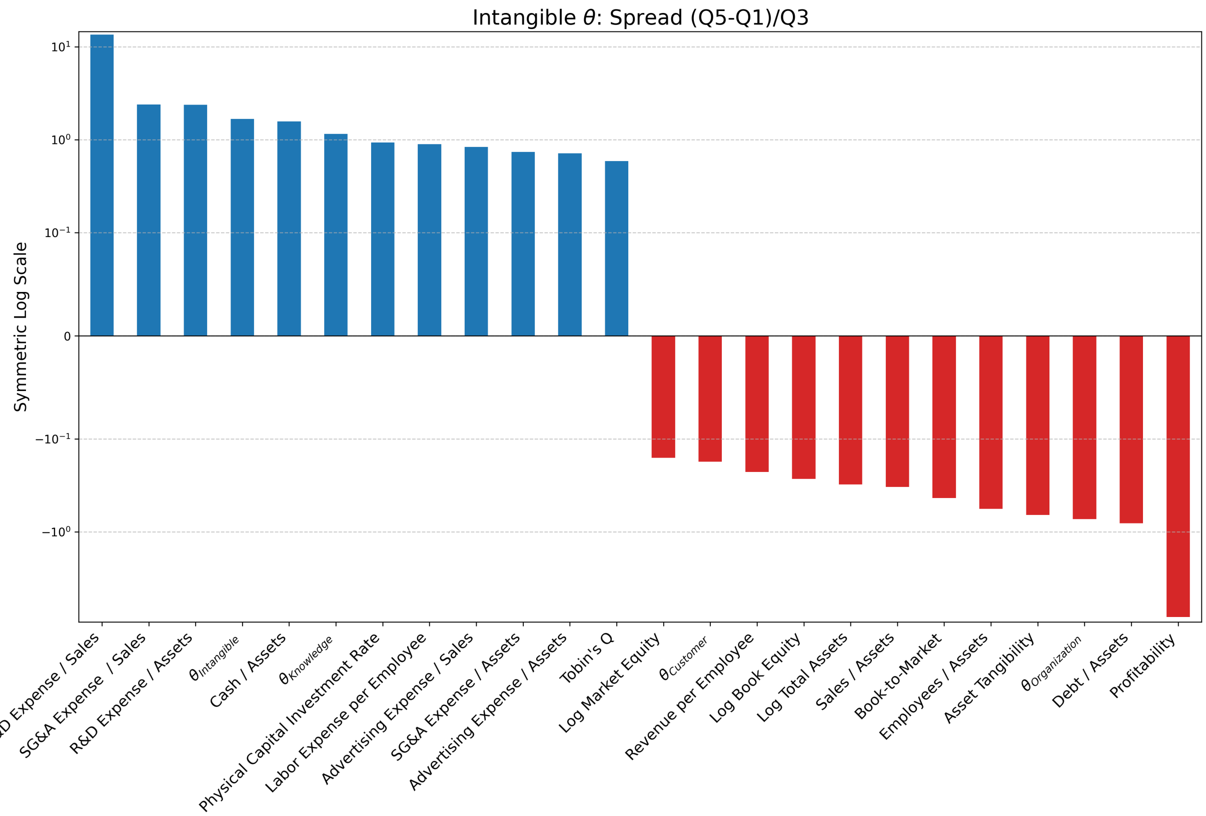


Figure 6: Q5-Q1 firm characteristics by θ

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

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 employee counts falling and labor expense per employee rising 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 higher θ^{int}, 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 **??** 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.

Table 4 Correlation Matrix of Intangible Capital Scores (Percentage)

	θ_{int}	θ_{know}	θ_{brand}	θ_{org}
θ_{int}	100.0%	67.9%	-24.4%	-47.6%
θ_{know}	67.9%	100.0%	-47.7%	-59.2%
θ_{brand}	-24.4%	-47.7%	100.0%	-42.6%
θ_{org}	-47.6%	-59.2%	-42.6%	100.0%

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.

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.[11]

These cases confirm that the θ^{int} components generally 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.

[11] 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	6.25	6.5	6.34	6.15	5.85
Log Book Equity	5.61	5.71	5.41	5.23	4.88
Log Total Assets	6.46	6.53	6.12	5.83	5.39
Log Intangible Capital	3.64	3.89	3.7	3.55	3.21
Intangible Capital Share	0.372	0.401	0.467	0.489	0.484
Knowledge Capital Share	0.335	0.335	0.334	0.327	0.311
Brand Capital Share	0.378	0.378	0.377	0.381	0.389
Book-to-Market	0.773	0.675	0.611	0.576	0.582
Tobin's Q	1.69	2	2.27	2.41	2.67
Profitability	0.122	0.0957	0.0463	-0.0329	-0.197
Asset Tangibility	0.261	0.243	0.208	0.188	0.173
Physical Capital Investment Rate	0.247	0.291	0.348	0.366	0.467
Sales / Assets	1.27	1.08	0.94	0.836	0.587
Cash / Assets	0.135	0.174	0.249	0.313	0.455
Debt / Assets	0.272	0.263	0.216	0.196	0.171
Employees / Assets	0.00797	0.00639	0.00485	0.00403	0.00301
Revenue per Employee	451	427	405	390	323
Labor Expense per Employee	74.4	69	75.5	102	145
Executive Compensation / Assets	3.44	3.58	4.45	4.82	4.71
R&D Expense / Assets	0.0142	0.0465	0.0907	0.139	0.258
Advertising Expense / Assets	0.0339	0.0312	0.0319	0.0278	0.0322
SG&A Expense / Assets	0.271	0.295	0.361	0.392	0.409
SG&A Expense / Sales	0.244	0.364	0.604	1.28	2.1
R&D Expense / Sales	0.0207	0.0857	0.276	2.51	14.3
Advertising Expense / Sales	0.0301	0.0332	0.0422	0.0341	0.0505
R&D Expense / SG&A Expense	0.0147	0.0633	0.144	0.197	0.262
Advertising Expense / SG&A Expense	0.111	0.16	0.211	0.157	0.0579
θ^{int}	0.106	0.133	0.172	0.197	0.264
θ^{know}	0.0416	0.149	0.254	0.373	0.568
θ^{cust}	0.463	0.455	0.421	0.358	0.229
θ^{org}	0.495	0.396	0.325	0.268	0.203

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. All firm characteristic variables are winsorized at the 1% level. The sample period is from 2002 to 2023.

Table 6 Characteristics by Brand Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Market Equity	6.01	6.29	6.3	6.27	6.22
Log Book Equity	5.08	5.42	5.43	5.43	5.47
Log Total Assets	5.66	6.08	6.15	6.16	6.28
Log Intangible Capital	3.21	3.62	3.7	3.71	3.75
Intangible Capital Share	0.439	0.435	0.444	0.454	0.429
Knowledge Capital Share	0.319	0.324	0.332	0.335	0.333
Brand Capital Share	0.388	0.383	0.378	0.377	0.378
Book-to-Market	0.596	0.618	0.611	0.644	0.743
Tobin's Q	2.47	2.28	2.19	2.16	1.95
Profitability	-0.135	-0.0264	0.0395	0.0574	0.0949
Asset Tangibility	0.21	0.213	0.215	0.21	0.223
Physical Capital Investment Rate	0.419	0.362	0.319	0.324	0.291
Sales / Assets	0.708	0.876	0.942	1.02	1.16
Cash / Assets	0.386	0.293	0.242	0.23	0.178
Debt / Assets	0.186	0.219	0.232	0.224	0.256
Employees / Assets	0.00522	0.00573	0.00499	0.00478	0.00489
Revenue per Employee	295	356	396	423	516
Labor Expense per Employee	105	97	76.3	62.1	66.7
Executive Compensation / Assets	3.98	4.09	4.44	4.38	3.79
R&D Expense / Assets	0.25	0.159	0.0977	0.0775	0.0454
Advertising Expense / Assets	0.0214	0.0232	0.0276	0.0354	0.0401
SG&A Expense / Assets	0.347	0.335	0.344	0.355	0.332
SG&A Expense / Sales	1.36	1.12	0.7	0.561	0.367
R&D Expense / Sales	14	4.93	0.747	0.223	0.0712
Advertising Expense / Sales	0.0241	0.0273	0.0344	0.0397	0.0392
R&D Expense / SG&A Expense	0.198	0.166	0.147	0.119	0.0551
Advertising Expense / SG&A Expense	0.0392	0.0822	0.157	0.21	0.207
θ^{int}	0.213	0.183	0.165	0.165	0.148
θ^{know}	0.423	0.344	0.278	0.223	0.123
θ^{cust}	0.137	0.273	0.382	0.482	0.647
θ^{org}	0.44	0.382	0.34	0.295	0.229

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. All firm characteristic variables are winsorized at the 1% level. 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	5.83	6.07	6.27	6.4	6.52
Log Book Equity	4.94	5.15	5.39	5.6	5.74
Log Total Assets	5.54	5.78	6.07	6.37	6.57
Log Intangible Capital	3.32	3.52	3.7	3.81	3.69
Intangible Capital Share	0.499	0.478	0.455	0.406	0.373
Knowledge Capital Share	0.325	0.324	0.33	0.331	0.335
Brand Capital Share	0.382	0.382	0.38	0.38	0.379
Book-to-Market	0.643	0.611	0.62	0.67	0.672
Tobin's Q	2.39	2.43	2.31	2.04	1.86
Profitability	-0.0866	-0.0508	0.00626	0.0656	0.107
Asset Tangibility	0.172	0.188	0.208	0.242	0.261
Physical Capital Investment Rate	0.391	0.395	0.335	0.288	0.275
Sales / Assets	0.787	0.825	0.933	1.02	1.13
Cash / Assets	0.369	0.335	0.27	0.201	0.156
Debt / Assets	0.186	0.203	0.222	0.244	0.262
Employees / Assets	0.00334	0.00391	0.00457	0.00566	0.00904
Revenue per Employee	416	384	400	415	374
Labor Expense per Employee	118	116	93.2	76.6	73.5
Executive Compensation / Assets	4.56	4.9	4.62	3.63	3.5
R&D Expense / Assets	0.189	0.163	0.122	0.0818	0.0429
Advertising Expense / Assets	0.0427	0.0315	0.0282	0.0295	0.0252
SG&A Expense / Assets	0.404	0.39	0.365	0.306	0.256
SG&A Expense / Sales	1.21	0.92	0.804	0.469	0.297
R&D Expense / Sales	6.16	4.21	1.56	0.516	0.157
Advertising Expense / Sales	0.0481	0.0367	0.033	0.0332	0.0248
R&D Expense / SG&A Expense	0.204	0.183	0.155	0.0975	0.0453
Advertising Expense / SG&A Expense	0.145	0.16	0.173	0.141	0.078
θ^{int}	0.246	0.209	0.175	0.139	0.106
θ^{know}	0.433	0.359	0.28	0.207	0.113
θ^{cust}	0.436	0.423	0.422	0.39	0.255
θ^{org}	0.131	0.218	0.298	0.402	0.631

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. All firm characteristic variables are winsorized at the 1% level. The sample period is from 2002 to 2023.

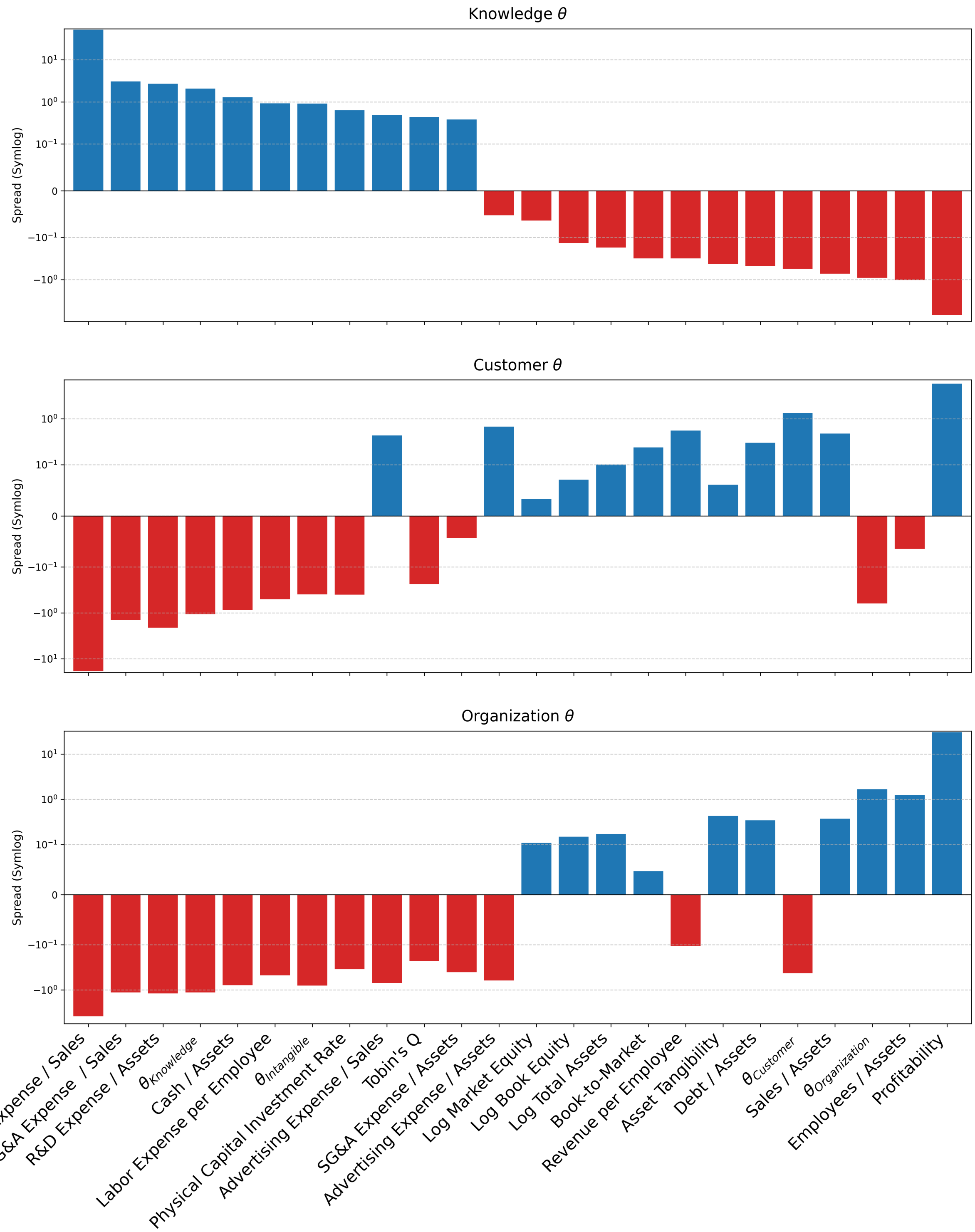


Figure 7: Q5-Q1 firm characteristics by θ

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 semantic theme scoring (STS), 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, the STS 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.

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

Belo, Frederico, Vito Gala, Juliana Salomao, and Maria Ana Vitorino, "Decomposing Firm Value," *Journal of Financial Economics*, 2022, *143* (2), 619–639.

Bybee, J. Leland, "The Ghost in the Machine: Generating Beliefs with Large Language Models," 2025. Working Paper.

Clayton, Christopher, Antonio Coppola, Matteo Maggiori, and Jesse Schreger, "Geoeconomic Pressure," NBER Working Paper 34020, National Bureau of Economic Research 2025.

Corrado, Carol, Charles Hulten, and Daniel Sichel, "Intangible Capital and U.S. Economic Growth," *Review of Income and Wealth*, 2009, *55* (3), 661–685.

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

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.

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

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.

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.

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.

Hoberg, Gerard and Gordon M. Phillips, "Scope, Scale, and Concentration: The 21st-Century Firm," *The Journal of Finance*, 2024, *80* (1), 415–466.

_ 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.

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

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

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.

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

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

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

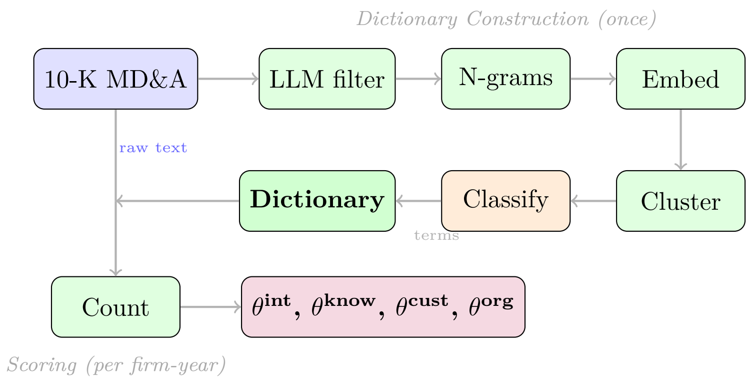
Sun, Qi and Mindy Z. Xiaolan, "Financing Intangible Capital," *Journal of Financial Economics*, 2019, *133* (3), 564–588.

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

Appendix A Additional Details on Methodology

Appendix A.1 Semantic Theme Scoring (STS) Methodology

This appendix provides additional details on the semantic theme scoring (STS) methodology. The goal is to construct a transparent, reproducible text scoring methodology to obtain firm-year measures of intangible investment intensity:



We separate the procedure into (i) an LLM-curated *dictionary construction* stage (Step 1–5) and (ii) a *scoring* stage where we calculate intensity measures for each firm-year document (Step 6).

Step 1: Extract SG&A-relevant text (LLM curation)

We start from raw *Item 7: Management Discussion and Analysis (MD&A)* text from 10-K filings.

1. Parse the 10-K, strip HTML and tables, and isolate Item 7.
2. Identify short context blocks around SG&A-related anchor terms (e.g. "SG&A", "selling and marketing", "general and administrative", "operating expenses").
3. For each context block and each expense component, call a lightweight open-source LLM whose sole job is to *extract* relevant sentences and return them in a structured format.

The quote-extraction prompt is:

```
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": [
  ]
}}
"""
"""
```

Note that the LLM is used only for text extraction, not classification. This is because the MD&A contains a large amount of boilerplate (liquidity discussion, accounting policies, generic risk factors). Using a small LLM as a *filter* allows us to focus subsequent steps on the subset of text that actually explains what the expense is, why the firm incurs it, and why it changes over time. Because the task is extraction rather than nuanced classification, a lightweight model is sufficient and inexpensive, and it is easy to swap in a different model without changing any of the downstream pipeline.

Step 2: Build an n-gram dictionary

From the extracted quotes, we build an interpretable universe of n-grams.

1. Tokenize each quote and extract bigrams and trigrams.

2. Restrict attention to noun-like phrases (e.g. "software development", "brand advertising expense") to avoid fragments that are unlikely to carry economic meaning.

3. Compute corpus-wide frequencies and keep the top N n-grams (in the slides, $N = 10{,}000$) as our "whitelist."

N-grams provide "atomic" semantic units: they are short enough to be interpretable and easy to label, yet rich enough to capture specific economic activities. Working at the n-gram level rather than with paragraph- or document-level embeddings allows us to construct a reusable dictionary that can be applied to new text without re-running any LLM calls. Context enhancements can still be introduced at the scoring stage, so the dictionary serves as a robust set of semantic anchors with potentially more flexible implementations.

Step 3: Embed n-grams and post-process the embeddings

We embed each n-gram in a high-dimensional semantic space using a modern sentence-transformer model (for example, `text-embedding-3-large`). This produces, for each n-gram w, an embedding vector $x_w \in \mathbb{R}^E$. Off-the-shelf embeddings are not directly suitable for clustering because of the well-known *anisotropy* problem: most vectors concentrate in a narrow cone, so cosine similarities are high even for semantically unrelated phrases. Our post-processing pipeline addresses this:

1. **PCA projection:** Project onto a lower-dimensional subspace spanned by the leading principal components, which removes noise dimensions and keeps the main semantic directions.

2. **Whitening (Optional):** Rescale each principal component so that it has unit variance; this spreads the cloud of points more evenly in all directions.

3. **Normalization:** Normalize each vector to unit length. After this step, Euclidean distance and cosine similarity are equivalent, which is convenient for K-means.

Raw embeddings often look like a narrow cone in $\mathbb{R}^E$: almost all angles are small, and pairwise cosine similarities are clustered. PCA and whitening round out this cone into something closer to a sphere, so that nearby phrases correspond to short Euclidean distances and cosine similarity becomes informative. A potential cost is that whitening downweights high-variance semantic directions, which may be important for some tasks. For our n-gram dictionary, however, this trade-off is attractive: we want clustering to reflect many semantic dimensions rather than a few dominant ones, and whitening prevents those directions from overwhelming the distance metric. The subsequent clustering step then identifies dense regions on this sphere that correspond to coherent economic concepts.

Step 4: Cluster n-grams into semantic communities

We next group the post-processed n-gram embeddings into communities of phrases that describe similar activities.

1. Run K-means clustering with a relatively large number of micro-clusters. This creates many fine-grained groups of semantically similar n-grams.

2. Construct a graph whose nodes are the micro-cluster centroids; connect nearby centroids with edges weighted by cosine similarity.

3. Apply a community-detection algorithm (Leiden) to this graph to merge micro-clusters into a few hundred "communities" of n-grams.

Ten thousand n-grams is too many to hand-label individually, while choosing an arbitrary K for K-means tends to give unstable, boundary-sensitive clusters. Our two-stage procedure involving

K-means micro-clusters followed by graph-based community detection lets the data determine how many economically meaningful communities there are, while keeping each community semantically tight (high within-community cosine similarity and low cross-community similarity). Because the resulting communities are relatively small and coherent, researchers can then overlay a researcher-driven classification (e.g., into knowledge, customer, and organization capital) tailored to the needs of the specific application.

Step 5: Classify communities into intangible categories

Each n-gram community is then classified into one of five mutually exclusive categories:

- **Knowledge capital**
- **Brand or customer capital**
- **Organization capital**
- **Non-intangible**

In practice, we classify at the community level (rather than n-gram by n-gram) by inspecting n-grams from each community and assigning the label that best describes the bulk of its content. This can be done by human coders or by a second, carefully prompted LLM, and it is a one-time cost. We treat a union of the set of all n-grams assigned to knowledge, brand or customer and organization capital as *intangible-investment*. The "whitelist" W is the set of the top-N n-grams, which we partition into three mutually exclusive categories: *intangible-investment*, and *non-intangible-investment*.

Classifying at the community level leverages the fact that neighboring n-grams in embedding space typically describe the similar underlying activities. This dramatically reduces the amount of manual labeling required and yields an economically interpretable taxonomy that can be reused in other applications.

Step 6: Score firm-year documents

Finally, we map the dictionary back to the raw MD&A text for each firm-year and construct frequency-based measures of intangible investment intensity.

For each firm-year:

1. Begin with the raw Item 7 MD&A text (not the LLM-extracted quotes).
2. Extract all bigrams and trigrams from the text. For each n-gram n, let $\text{cnt}(n)$ denote its frequency in the document.
3. For each community k, compute the community count by summing over its constituent n-grams:
$$C_k = \sum_{n \in k} \text{cnt}(n).$$

4. Let $C_{\text{white}} = \sum_k C_k$ denote the total count across all whitelisted communities. The document-level intangible share is then:
$$\theta^{\text{int}} = \frac{\sum_{k \in \mathcal{I}} C_k}{C_{\text{white}}},$$
where $\mathcal{I}$ is the set of communities classified as intangible-investment.

5. Category-specific shares θ^{know}, θ^{cust}, and θ^{org} are defined analogously, substituting the relevant subset of communities in the numerator.

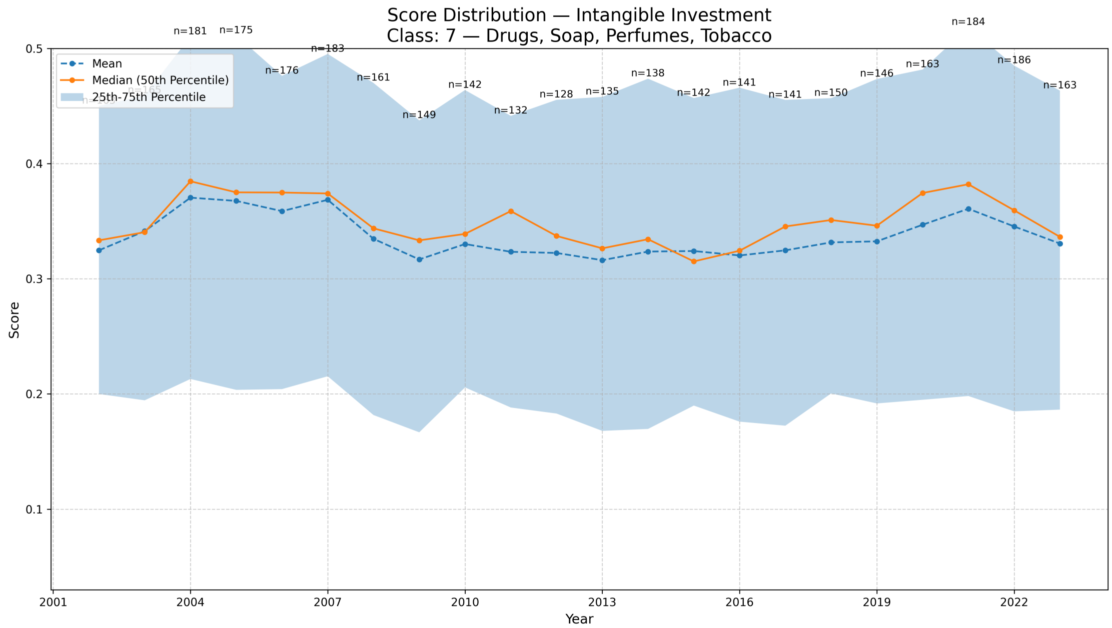
Two design choices are important here:

- *Score on raw text.* We deliberately score n-grams on the raw MD&A, not on the LLM-extracted quotes. The extracted quotes are by construction focused on SG&A, which would mechanically inflate the measured share of intangible-related discussion. Raw text provides a fair representation of what the firm emphasizes overall.

- *Restrict the denominator to the whitelist.* By using only n-grams in W in the denominator, we ensure an apples-to-apples comparison across firms. Rare or idiosyncratic n-grams that never enter the top-N dictionary simply drop out of both the numerator and the denominator. This raises a natural concern that we might miss firm-specific yet economically important n-grams. However, if an idiosyncratic trigram expresses a concept that is truly important in a broader sense (e.g., an intangible-investment concept), then at least one of its constituent bigrams is, by construction, likely to appear frequently enough across firms to enter the whitelist. In that case, the underlying economic concept is still captured through the corresponding bigram, even if the firm-specific trigram itself is excluded.

The resulting $\theta^{\text{int},d}$ is a transparent, easily replicable proxy for the fraction of SG&A-related discussion devoted to activities that resemble intangible investment. In the main text, we show that this text-based measure behaves in a way consistent with economic priors and helps explain cross-sectional and time-series variation in firm outcomes.

Appendix B Additional Figures and Tables



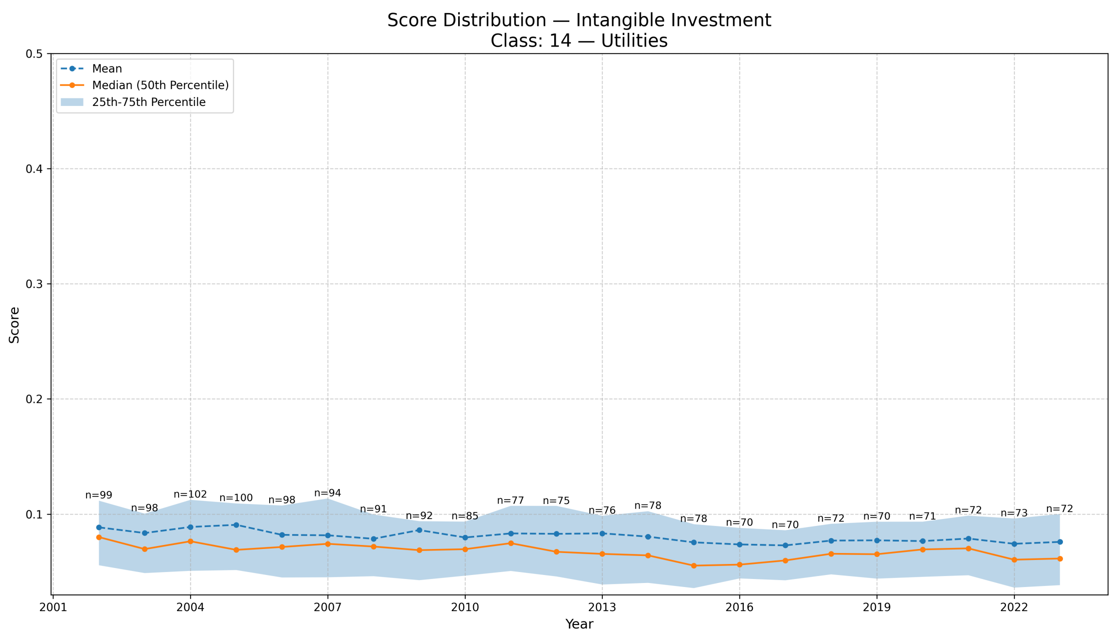


Figure B.1: Intangible Intensity Distribution for Select Industries

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**
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
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

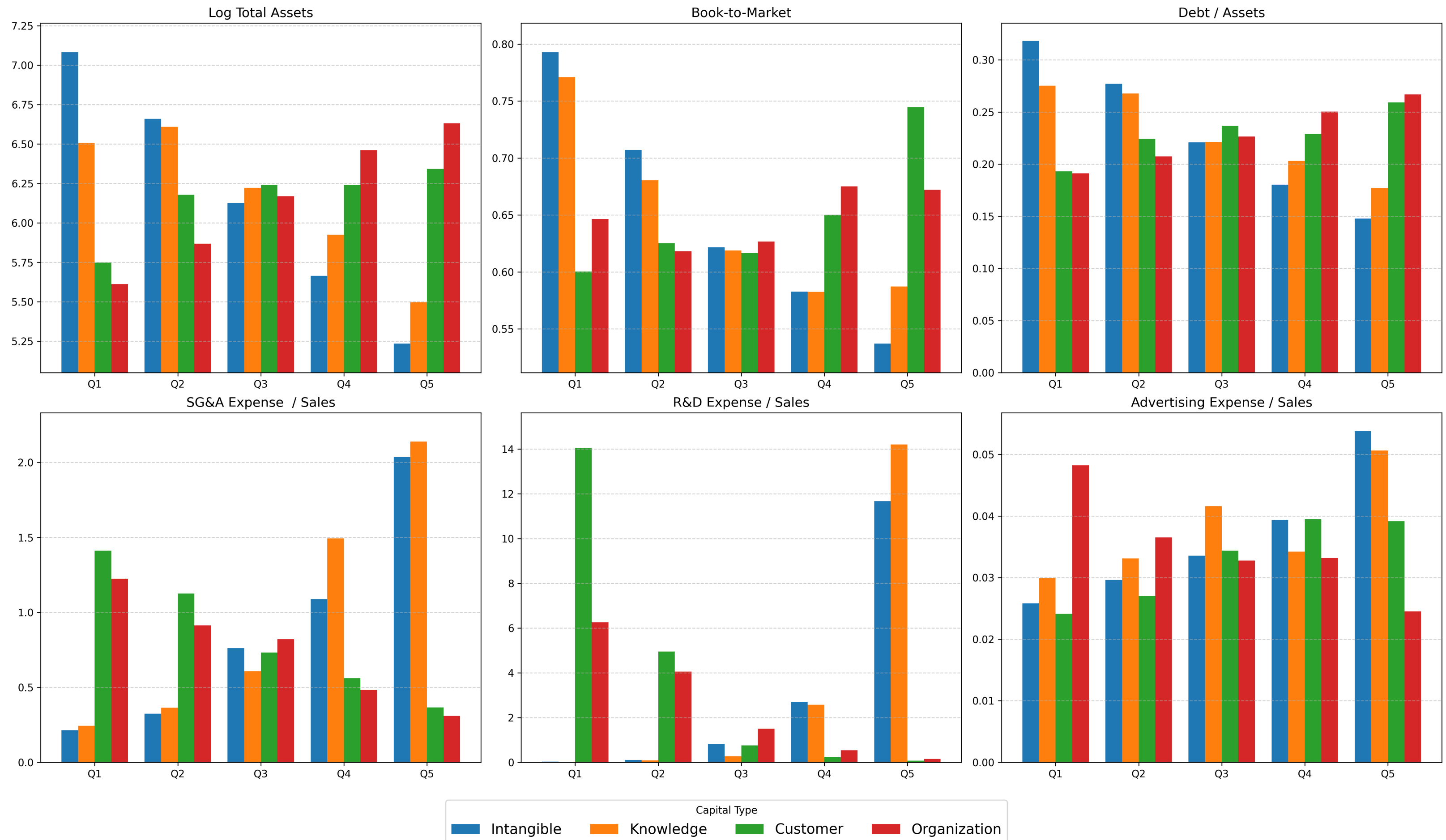


Figure B.2: Time-Averaged Values Across Quintiles

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

Table B.2 Summary Statistics (Summary By Industry Ff17)

Industry	Type	Mean	SD	P25	P50	P75
Food	θ^{int}	0.129	0.061	0.086	0.116	0.159
	θ^{cust}	0.540	0.174	0.420	0.549	0.669
	θ^{know}	0.120	0.114	0.038	0.090	0.167
	θ^{org}	0.340	0.169	0.213	0.312	0.438
Mining and Minerals	θ^{int}	0.086	0.043	0.054	0.075	0.108
	θ^{cust}	0.410	0.215	0.245	0.396	0.571
	θ^{know}	0.132	0.124	0.037	0.103	0.194
	θ^{org}	0.458	0.213	0.290	0.437	0.608
Oil and Petroleum Products	θ^{int}	0.058	0.030	0.038	0.053	0.071
	θ^{cust}	0.304	0.184	0.167	0.286	0.412
	θ^{know}	0.189	0.145	0.077	0.164	0.276
	θ^{org}	0.507	0.193	0.368	0.500	0.639
Textiles, Apparel & Footwear	θ^{int}	0.114	0.053	0.078	0.103	0.139
	θ^{cust}	0.587	0.157	0.500	0.597	0.690
	θ^{know}	0.093	0.093	0.033	0.070	0.124
	θ^{org}	0.320	0.145	0.217	0.302	0.400
Consumer Durables	θ^{int}	0.139	0.075	0.085	0.120	0.176
	θ^{cust}	0.481	0.178	0.360	0.485	0.610
	θ^{know}	0.215	0.156	0.087	0.190	0.324
	θ^{org}	0.304	0.170	0.179	0.272	0.398
Chemicals	θ^{int}	0.111	0.079	0.061	0.087	0.128
	θ^{cust}	0.419	0.188	0.280	0.406	0.556
	θ^{know}	0.238	0.184	0.083	0.200	0.371
	θ^{org}	0.343	0.176	0.224	0.309	0.429
Drugs, Soap, Perfumes, Tobacco	θ^{int}	0.286	0.134	0.175	0.286	0.389
	θ^{cust}	0.299	0.201	0.126	0.265	0.436
	θ^{know}	0.503	0.236	0.323	0.551	0.709
	θ^{org}	0.198	0.123	0.121	0.167	0.243
Construction and Construction Materials	θ^{int}	0.106	0.059	0.065	0.094	0.134
	θ^{cust}	0.492	0.214	0.333	0.500	0.658
	θ^{know}	0.108	0.108	0.037	0.078	0.146
	θ^{org}	0.400	0.217	0.229	0.370	0.544
Steel Works Etc	θ^{int}	0.081	0.042	0.051	0.073	0.104
	θ^{cust}	0.432	0.205	0.307	0.444	0.566
	θ^{know}	0.134	0.115	0.037	0.119	0.203
	θ^{org}	0.434	0.213	0.290	0.400	0.529
Fabricated Products	θ^{int}	0.102	0.054	0.069	0.090	0.119
	θ^{cust}	0.462	0.162	0.376	0.469	0.560
	θ^{know}	0.174	0.128	0.087	0.149	0.237
	θ^{org}	0.364	0.178	0.233	0.333	0.457

Continued on next page

Continued from previous page

Industry	Type	Mean	SD	P25	P50	P75
Machinery and Business Equipment	θ^{int}	0.168	0.087	0.102	0.150	0.222
	θ^{cust}	0.388	0.154	0.280	0.379	0.485
	θ^{know}	0.339	0.176	0.212	0.349	0.463
	θ^{org}	0.273	0.140	0.174	0.248	0.344
Automobiles	θ^{int}	0.118	0.057	0.081	0.104	0.146
	θ^{cust}	0.400	0.175	0.267	0.384	0.531
	θ^{know}	0.252	0.168	0.111	0.234	0.373
	θ^{org}	0.349	0.165	0.235	0.324	0.443
Transportation	θ^{int}	0.104	0.056	0.063	0.092	0.136
	θ^{cust}	0.335	0.184	0.195	0.318	0.471
	θ^{know}	0.129	0.147	0.027	0.076	0.174
	θ^{org}	0.537	0.234	0.333	0.553	0.714
Utilities	θ^{int}	0.077	0.051	0.051	0.067	0.087
	θ^{cust}	0.472	0.172	0.376	0.483	0.580
	θ^{know}	0.137	0.096	0.065	0.121	0.194
	θ^{org}	0.391	0.170	0.278	0.367	0.471
Retail Stores	θ^{int}	0.114	0.060	0.075	0.101	0.135
	θ^{cust}	0.541	0.171	0.438	0.555	0.658
	θ^{know}	0.071	0.073	0.019	0.050	0.101
	θ^{org}	0.388	0.176	0.265	0.368	0.485
Other	θ^{int}	0.204	0.125	0.102	0.179	0.282
	θ^{cust}	0.357	0.194	0.200	0.362	0.488
	θ^{know}	0.310	0.237	0.105	0.273	0.459
	θ^{org}	0.333	0.210	0.175	0.276	0.438

Table B.3 Variable Definitions and Formulas

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

Notes: k_{int} is calculated using the perpetual inventory method and capitalizes 100% of xsga with 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 IPO date. 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	Industry	SG&A/Sales	Median	R&D/Sales	Median	Adv/Sales	Median	θ^{int}	θ^{know}	θ^{cust}	θ^{org}
Apple Inc.	AAPL	Other	12.0%	47.4%	6.0%	20.8%		1.9%	1	4	3	3
Microsoft	MSFT	Other	27.3%	47.4%	12.3%	20.8%	0.9%	1.9%	4	3	5	1
Alphabet	GOOG	Other	26.4%	47.4%	12.3%	20.8%	3.1%	1.9%	2	2	5	3
Amazon	AMZN	Rtail	36.7%	24.4%	11.9%	0.0%	3.6%	2.2%	4	4	4	2
Tesla	TSLA	Cars	13.2%	14.5%	4.8%	3.7%		0.9%	4	4	3	2
Nvidia	NVDA	Machn	27.6%	26.3%	19.6%	7.7%		0.6%	3	5	4	1
Meta Platforms	META	Other	41.1%	47.4%	20.9%	20.8%	2.5%	1.9%	3	4	4	2
Procter & Gamble	PG	Cnsum	27.4%	52.7%	2.5%	25.7%	10.8%	4.1%	1	1	3	5
Walmart	WMT	Rtail	20.2%	24.4%	0.0%	0.0%	0.7%	2.2%	1	3	1	1
Eli Lilly & Co.	LLY	Cnsum	47.5%	52.7%	27.9%	25.7%	4.2%	4.1%	2	5	2	1
Pfizer Inc.	PFE	Cnsum	29.3%	52.7%	17.0%	25.7%	2.5%	4.1%	2	3	4	1
Home Depot	HD	Cnstr	16.6%	13.8%	0.0%	0.8%	0.7%	0.5%	1	2	1	5
Coca-Cola	KO	Food	31.0%	21.1%		0.4%	10.6%	2.8%	4	4	2	4
AbbVie	ABBV	Other	33.8%	47.4%	14.3%	20.8%	3.7%	1.9%	2	4	2	2
Merck & Co.	MRK	Cnsum	43.8%	52.7%	25.1%	25.7%	4.1%	4.1%	2	2	4	4
PepsiCo	PEP	Food	39.1%	21.1%	0.9%	0.4%	4.4%	2.8%	4	5	3	3
Verizon	VZ	Other	22.0%	47.4%		20.8%	2.5%	1.9%	2	2	5	4
Thermo Fisher	TMO	Machn	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 (lowest) to 5 (highest). Median represents the industry median value for the preceding column.