Intangible Intensity

Andrea L. Eisfeldt* Barney Hartman-Glaser†

Edward T. Kim‡ Ki Beom Lee§

July 2026

LINK TO GITHUB CODE REPOSITORY

Abstract

We develop a text-based measure of intangible investment intensity derived from 10-K filings, and offer a general methodology for semantic theme scoring (STS). The approach further decomposes disclosure text into knowledge, customer, and organization capital. Knowledge-intensive firms are small, young, and R&D-driven, customer-intensive firms are profitable and commercially oriented, and organization-intensive firms are large incumbents built on operational scale. We calibrate the scores to observed spending and accumulate firm-level intangible capital stocks, which more than triple the explanatory power of a franchise-value benchmark for market-to-book ratios. Firms with more intangible capital also borrow less and hold substantially more cash. Managerial language thus embeds forward-looking signals about intangible investment that accounting data obscure.

Keywords: Intangible capital, textual analysis, corporate disclosure, corporate valuation

JEL classification: E22, G32, C81, D22

For helpful comments, the authors thank Ali Kakhbod, Indrajit Mitra, as well as seminar and conference participants at UCLA Anderson, Michigan Ross, NBER Digital Economics and AI Tutorial, SFS Cavalcade, and FIRS Conference. *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 300 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.

We then show that the text-based scores track observed spending. Wherever firms disclose disaggregated intangible expenditure, industry-level θ scores line up closely with actual spending shares. We use this relation to calibrate the scores, estimating one multiplier per component on the pooled sample of firms with observable spending. Multipliers estimated within broad industry groups are similar and leave the results unchanged. Accumulating the calibrated scores with the perpetual inventory method yields dollar-denominated stocks of knowledge, customer, and

organization capital for the full universe of public firms.

We use the decomposed stocks first to explain why market-to-book ratios deviate from one. The book value of equity omits internally generated intangible capital, and market values embed franchise value. Measuring both, we find that the calibrated intangible stocks, together with a simple franchise-value proxy, explain roughly 40% of the within-year cross-sectional variation in market-to-book, with intangible capital alone accounting for more than a quarter. The market prices the split between knowledge and organization capital, a distinction that aggregate intangible measures cannot see and that our firm-level decomposition identifies. The same composition matters for financing: in standard capital-structure regressions, the calibrated intangible stock is a robust cross-sectional correlate of lower leverage and substantially higher cash holdings, with the cash association again dominated by knowledge capital.

Overall, our paper provides a scalable and interpretable measure of intangible investment intensity from textual disclosures, available wherever firms file 10-Ks and independent of what firms choose to report as R&D or advertising. The measure and its components speak directly to questions of firm growth, valuation, and capital structure, and they remain informative where accounting data are incomplete or silent. Our contribution is a measurement tool that complements, rather than replaces, accounting-based approaches to intangible capital.

Our research contributes to multiple strands of the literature. The first concerns measuring intangible capital from accounting data. 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 obtained by Hulten and Hao (2008), who apply aggregate intangible-investment shares from Corrado et al. (2009) to a composite of six large pharmaceutical firms (Zhang, 2013; Eisfeldt and Papanikolaou, 2014; Peters and Taylor, 2017; Sun and Xiaolan, 2019; Falato et al., 2022; Chan et al., 2001). 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

[1]See also Ewens et al. (2025), who use data on firms that have undergone mergers and acquisitions in related work.

(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 67% and 43% of firm-years in our sample, our intangible-intensity score covers 84% (92% after within-firm interpolation). Second, the score measures the intensity of three types of intangible capital: knowledge, customer, and organization capital. In the absence of explicit reporting of SG&A subcomponents, our score helps shed new light on these crucial intangible assets.

A second strand of work studies the role of intangibles in firm valuation and financial disclosure. The declining value-relevance of accounting data has been attributed in part to the inadequate treatment of intangibles (Lev and Zarowin, 1999; Daniel and Titman, 2006), and early work shows that capitalizing R&D and software costs produces measures with significant explanatory power for firm value (Lev and Sougiannis, 1996; Aboody and Lev, 1998).[2] More recently, Filipovic and Wagner (2026) and David et al. (2023) use keyword dictionaries to measure intangible-related language in M&A announcements and 10-K filings, respectively, finding that textual intangible signals predict deal outcomes and stock returns. Our approach covers the full universe of public firms, constructs its dictionary endogenously from the corpus rather than from a pre-specified word list, and offers a general methodology applicable to any text classification task.

Finally, our work relates to the growing use of alternative data in finance. Text data and the use of LLMs have become increasingly popular (Gentzkow and Shapiro, 2010; Loughran and McDonald, 2011; Baker et al., 2016; Hoberg and Phillips, 2016, 2025; 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 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

[2]At the macroeconomic level, Corrado et al. (2005) and Corrado et al. (2009) introduce a comprehensive framework for measuring intangible investment in national accounts and show that including intangibles substantially raises measured U.S. business investment.

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.[3]

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 converts the intensity scores into intangible capital stocks and calibrates them to observed spending shares. Section 6 uses the calibrated stocks to explain the cross-section of market-to-book ratios. Section 7 relates intangible capital to leverage and cash holdings. Section 8 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 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).[4] 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 selling and marketing (customer capital) are disclosed unevenly across firms, a pattern

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

[4]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.

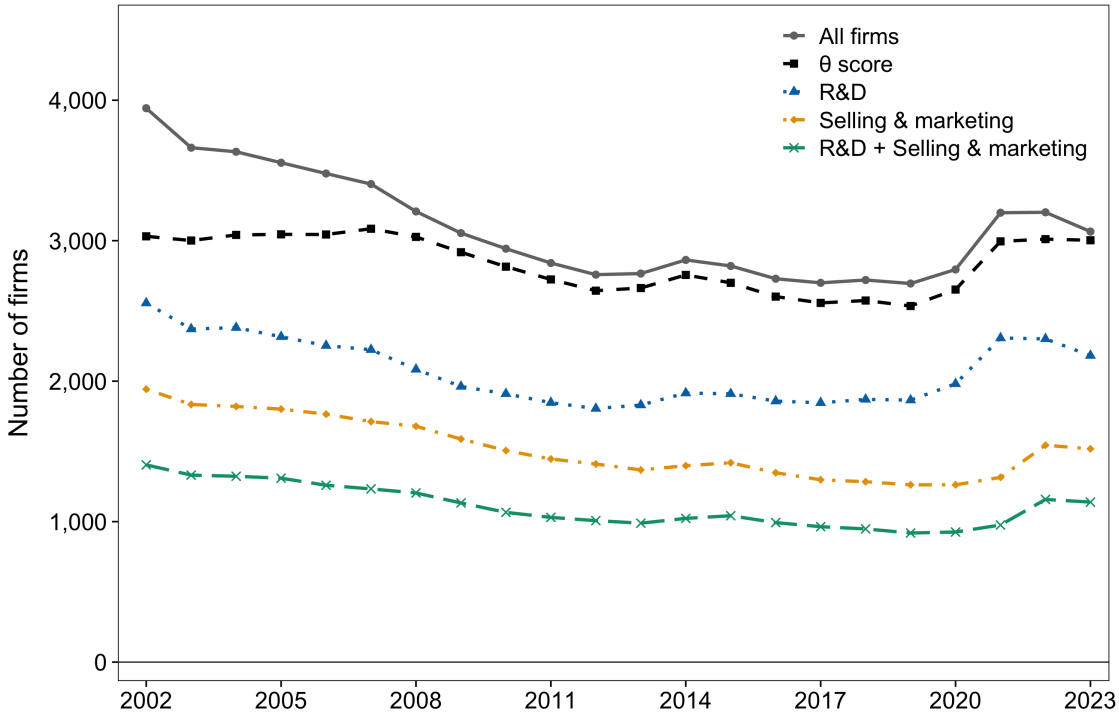


Figure 1: Coverage of Spending Variables

Notes: This figure shows, by fiscal year, the number of publicly traded firms in the sample, the number with a text-based θ score, and the number reporting R&D and selling and marketing expense items. R&D is from Compustat and selling and marketing is from Capital IQ, measured as the consolidated selling-and-marketing item where reported, the marketing item for firms without it, and the advertising item for firms without either, with items reported as zero or negative treated as missing. The bottom series counts firms reporting both R&D and selling and marketing.

that is clearly visible in Figure 1, while investments towards organization capital are not reported as a standalone item at all. As of 2023, reporting rates for R&D and selling and marketing are roughly 71% and 50%, respectively, and just over 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.[5] 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 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 LLM, 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

[5]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`.

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 standard cleaning procedures and create a frequency-ranked "master list." Our analysis focuses on the top $N = 30{,}000$ n-grams, or "whitelist", which balances coverage and tractability.

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 a large number of fine-grained micro-clusters. We treat the micro-clusters themselves as nodes in a graph whose edges are formed by aggregating mutual nearest-neighbor relations among their member n-grams, and apply the Leiden community detection algorithm to identify coherent, higher-level communities of semantically related n-grams. This process yields roughly 300 communities. Each community is then classified by LLM, with researcher review, into one of five categories: knowledge, customer, organization, non-intangible and unknown. The Unknown category prevents the LLM from being forced into a binary assignment when a community is genuinely ambiguous, reducing hallucinated classifications. Combining algorithmic community identification and clustering with LLM classification under researcher review effectively leverages the scalability enabled by modern LLMs with the domain knowledge of the researcher. With the small number of communities, the researcher can directly inspect every assignment. 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 $N = 30{,}000$ 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 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.

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

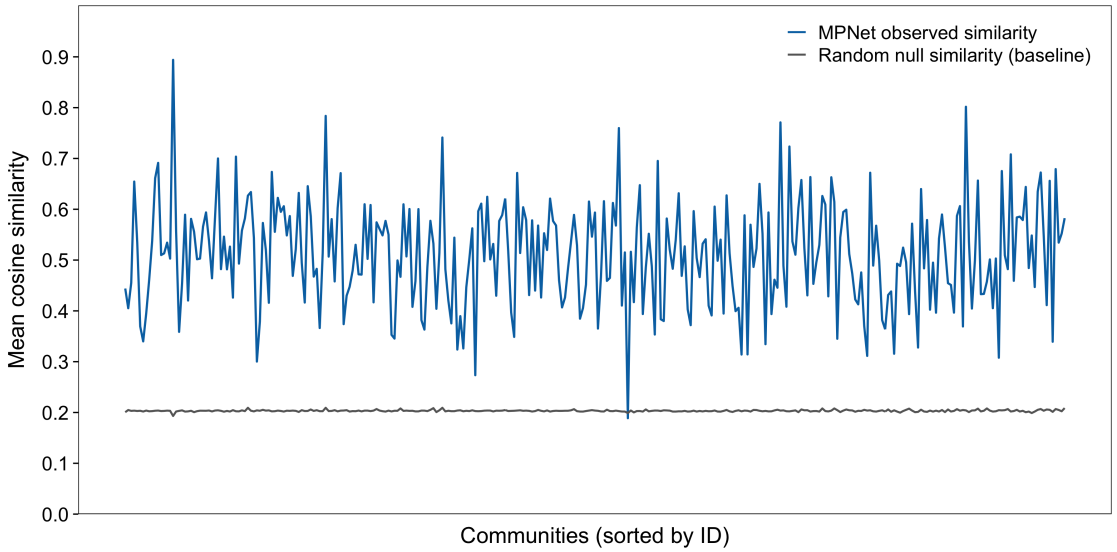


Figure 2: Community Cohesion Benchmarking

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

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 $N = 30{,}000$ 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.

[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.

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. We adopt $N = 30{,}000$ as our baseline (robust to larger samples), which offers a good balance between coverage and tractability.

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

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

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

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

   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. Our panel contains 68,032 firm-year observations of Compustat-CRSP matched firms from 2002 to 2023, roughly 3,100 per year. We construct θ^{int} directly from 10-K filings for 57,249 of these firm-year observations (84%). 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 62,423 firm-year observations, or 92% of the panel; the remaining firm-years belong to firms with no scorable 10-K text in any year.

Table 1 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 0.16 and the corresponding average for the non-intangible share is 0.84, 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 0.09 to 0.22, indicating substantial dispersion across firms. Decomposing θ^{int} into its subcategories reveals that organization capital accounts for the largest fraction of the score, followed by knowledge capital, with customer capital contributing the smallest share, consistent with the structure of SG&A expenses, which are more heavily weighted toward operational and administrative activities than toward customer-facing spending.

Figure 3 summarizes the evolution of θ^{int} over time. The median score is stable, moving within a narrow band around 0.13 to 0.16 throughout 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,

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

Table 1: Distribution of θ Scores and Subcomponents

Type	Mean	SD	P25	P50	P75
θ^{int}	0.163	0.095	0.090	0.138	0.219
θ^{know}	0.057	0.067	0.013	0.031	0.073
θ^{cust}	0.042	0.030	0.021	0.035	0.055
θ^{org}	0.065	0.037	0.037	0.057	0.085
θ^{nint}	0.837	0.095	0.781	0.862	0.910

Notes: This table reports the mean, standard deviation, and quartiles of θ^{int}, its three subcomponents (θ^{know}, θ^{cust}, θ^{org}), and the non-intangible share (θ^{nint}) across the 62,423 firm-year observations with non-missing interpolated scores, 2002–2023. The subcomponents sum to θ^{int}, and $\theta^{int} + \theta^{nint} = 1$.

the wide gap between the 10th and 90th percentiles 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 panel (b) of Figure 3, the median ratio of SG&A to total assets drifts down from roughly 29% in 2002 to 21% in 2020 and then rebounds to 27% by 2023, while the 90th percentile widens sharply to 0.95 at the end of 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 2022 and 2023, the two series diverge, with SG&A spending intensity rising 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 3 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 4, by plotting intensity measures across the Fama–French 17

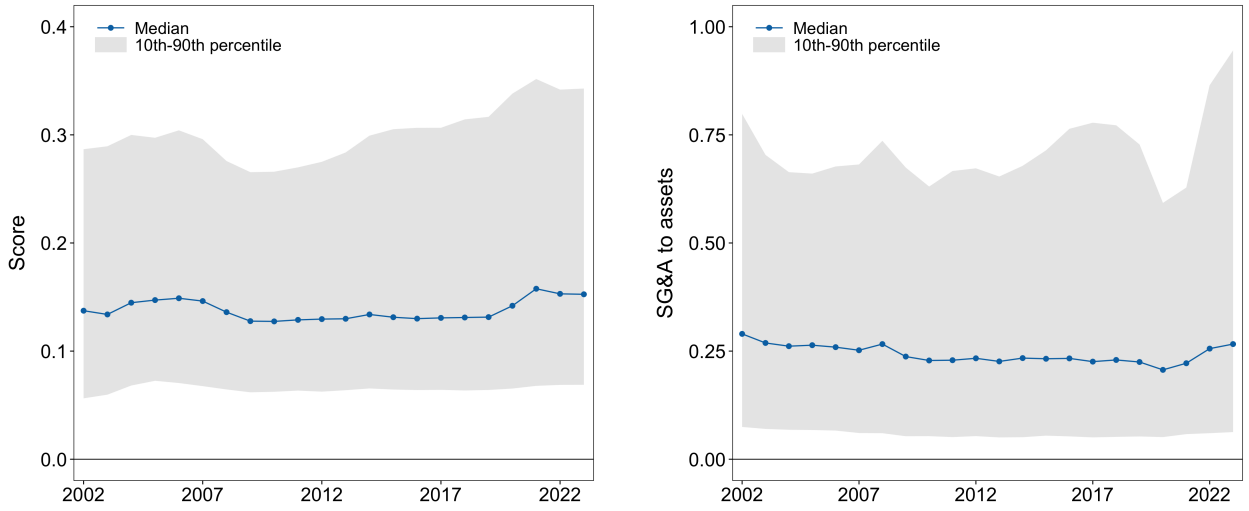
industry classification, slightly modified to better reflect modern firm classifications,[9][10] The average intangible intensity of roughly 0.26 for the "Healthcare" industry is the highest of the sample, more than one and a half times the aggregate average of 0.16, and is knowledge-dominated (0.133 knowledge, 0.041 customer, 0.085 organization), consistent with the prevalence of biotechnology and life-sciences firms.[11] In contrast, industries with the lowest average intangible intensity include "Steel Works," "Oil and Petroleum Products," and "Mining and Minerals," among others, consistent with their dependence on tangible capital and established production processes. At the other end of the spectrum, "Technology and Communications" displays the second-highest intensity at roughly 0.19, followed by the residual "Other" industry and "Fabricated Products and Machinery," both near 0.13 and containing many high technology and business services firms; "Travel and Entertainment" sits mid-pack at roughly 0.11 and is comparatively organization-heavy. 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 "Healthcare" 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" and is comparatively high within "Technology and Communications," consistent with the increasing importance of software and digital platforms. Customer capital is most prominent in "Technology and Communications," with the "Food and Consumer Staples" and "Textiles, Apparel, and Footwear" industries also ranking highly, reflecting the centrality of consumer-facing differentiation. In contrast, organization capital plays an outsized role, as a share of industry intensity, 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 within "Travel and Entertainment," where complementarities with human capital and operational scale are particularly salient. These component-level differences reinforce

[9]The modification reassigns a small set of SIC codes, following the adjustments of our thirteen-industry classification. For example, prepackaged software (SIC 7372) and telephone communications (4813) move from Other into Technology and Communications, surgical and medical instruments (3841) and hospitals (8062) move into Healthcare, and hotels (7011) move into Travel and Entertainment.

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

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



(a) Intangible Intensity Score

(b) SG&A to Total Assets

Figure 3: Distribution of Intangible Intensity Over Time

Notes: This figure reports, by fiscal year, the median and the 10th to 90th percentile range of θ^{int} across all firm-years (panel a), together with the median and the 10th to 90th percentile range of the ratio of SG&A expense to total assets (panel b).

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.

As a further check on this measure, we benchmark STS against an alternative in which a frontier LLM is prompted directly to classify each firm's intangible investment intensity from the same Item 7 text. The two approaches agree on the underlying economic signal: across all models and intangible capital components, STS θ scores rise monotonically with the LLM-assigned rating. Direct LLM scoring, however, carries substantial intrinsic classification noise. Within-model disagreement across repeated runs of the same model on identical documents (approximately 26%) is statistically indistinguishable from cross-model disagreement between different LLMs (approximately 29–30%), indicating that most disagreement reflects noise in the classification task itself rather than genuine differences across models; organization capital is the noisiest component, and setting the sampling temperature to zero yields only a modest improvement. Direct LLM

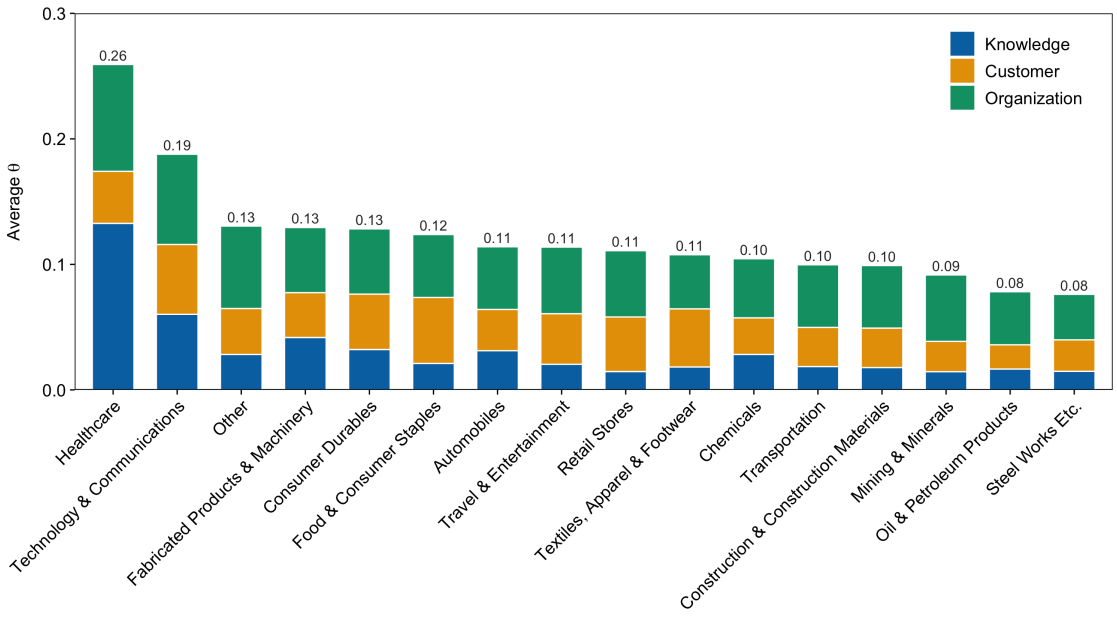


Figure 4: Intangible Intensity and Its Components Across Industries

Notes: This figure reports 2002–2023 time-averaged mean raw component scores (knowledge, customer, organization) by industry; components stack to the total average raw θ^{int}. Industries are sorted in descending order of total average raw θ^{int}. Utilities and Financials are absent from the estimation sample.

scoring is also substantially more costly and, absent an aggregation procedure across multiple runs, less reproducible than our approach. STS instead front-loads the semantic work into dictionary construction, which keeps the scoring step itself deterministic, fast, and auditable, an advantage for reproducibility and scale.[12]

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 industry classification introduced above as our baseline.

[12]See Appendix A.3 for the full comparison, including model-by-model scatter plots and disagreement rates.

Table 2: Characteristics by Intangible Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Total Assets	6.93	6.46	6.04	5.71	5.27
Tobin's Q	2.37	2.91	3.13	3.21	3.54
Book-to-Market	0.671	0.603	0.565	0.517	0.452
Profitability	0.118	0.0622	0.0108	-0.0378	-0.116
Asset Tangibility	0.29	0.25	0.214	0.187	0.148
Physical Capital Investment Rate	0.211	0.269	0.327	0.403	0.476
R&D Expense / Sales	0.0422	0.666	1.98	5	6.38
Advertising Expense / Sales	0.0206	0.0282	0.032	0.0394	0.0526
SG&A Expense / Sales	0.231	0.481	0.866	0.86	1.43
Cash / Assets	0.117	0.184	0.262	0.331	0.417
Debt / Assets	0.309	0.275	0.229	0.194	0.146
Employees / Assets	0.00634	0.00534	0.00497	0.00476	0.00375
Labor Expense per Employee	60.1	84.9	94.9	98.5	117
θ^{int}	0.0778	0.124	0.162	0.201	0.269
θ^{know}	0.0174	0.0347	0.054	0.0759	0.11
θ^{cust}	0.0243	0.036	0.0421	0.0473	0.0623
θ^{org}	0.0358	0.0532	0.0651	0.0772	0.0949

Notes: This table presents the equal-weighted average of firm-level statistics for portfolios sorted on θ^{int}. All sorts are done within industries. All firm characteristic variables are winsorized at the 1% level. The sample period is from 2002 to 2023. The full set of firm characteristics is reported in the appendix version of this table.

4.1.1 Total Intangible Intensity, θ^{int}

We begin by examining average firm characteristics sorted on θ^{int}. Table 2 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) in terms of size, as measured by total assets. Despite their smaller size, these high intensity firms command sharply higher valuations: the average Tobin's Q rises from 2.37 for Q1 to 3.54 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 0.118 to –0.116 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

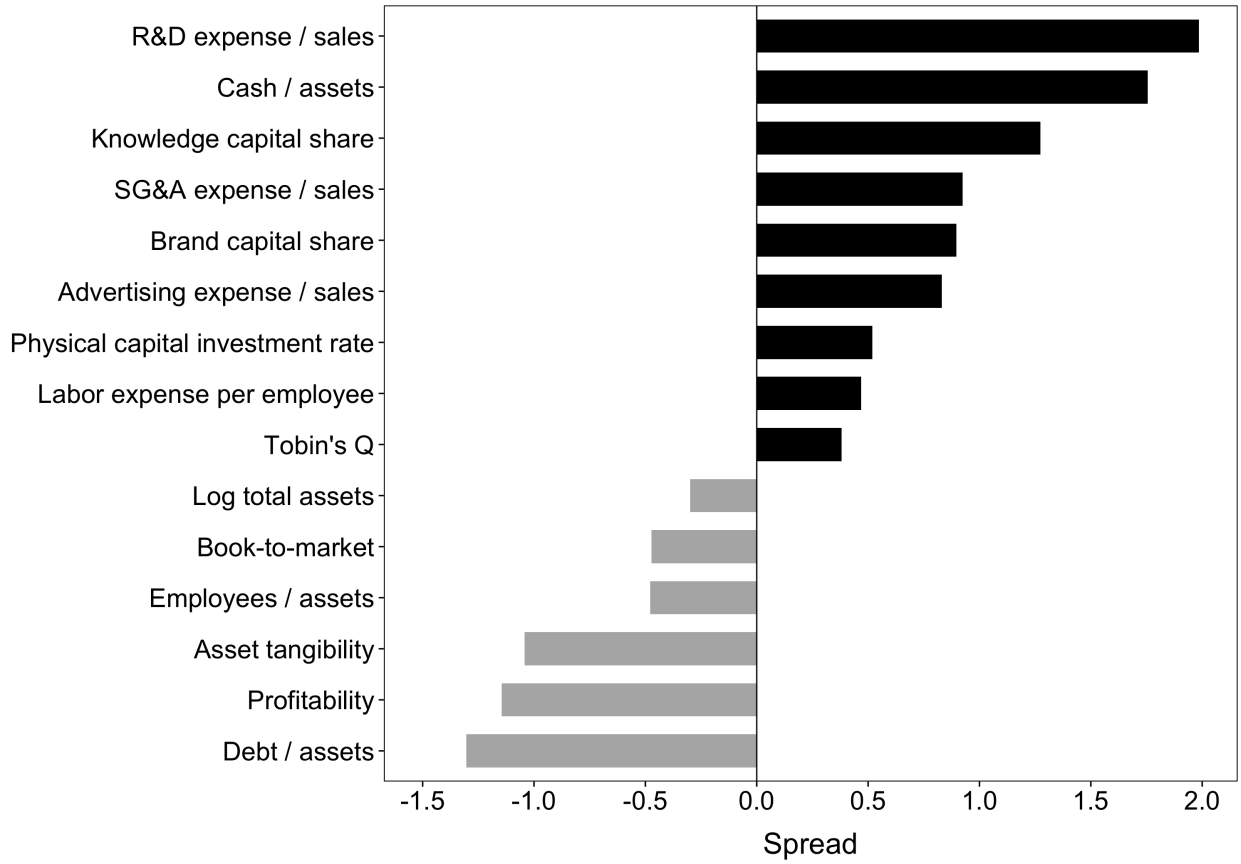


Figure 5: Q5-Q1 firm characteristics by θ

Notes: This figure plots the difference in median firm characteristics between the top and bottom quintiles of firms sorted on θ^{int} within industry-year cells, scaled by the absolute median of the middle quintile, that is $(\text{Q5}_{med} - \text{Q1}_{med})/|\text{Q3}_{med}|$; characteristics are winsorized at the 1% level. The variable set matches Table 2.

aligns with their escalating intangible and operational expenditure intensity: the R&D expense to sales ratio rises from 0.042 to 6.38, and the SG&A expense to sales ratio increases from 0.23 to 1.43. Physical investment rates also expand monotonically across quintiles (0.211 for Q1 vs. 0.476 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 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 60.1 for Q1 to 117 for Q5. 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.

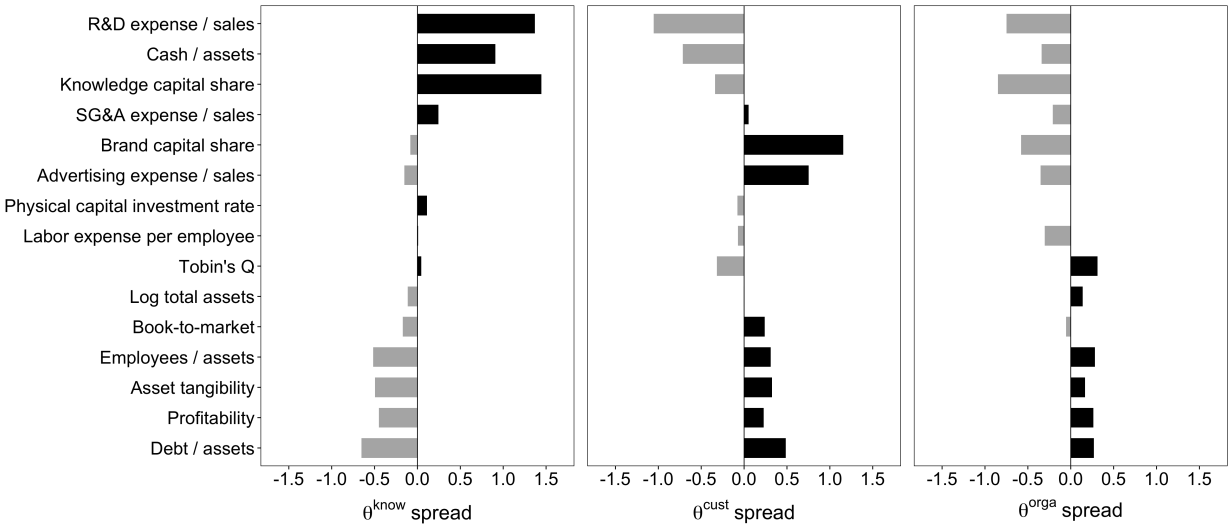


Figure 6: Q5-Q1 firm characteristics by θ component

Notes: This figure plots the difference in median firm characteristics between the top and bottom quintiles of firms sorted on θ^{know} (left panel), θ^{cust} (middle panel), and θ^{org} (right panel) within industry-year cells, scaled by the absolute median of the middle quintile; characteristics are winsorized at the 1% level. The variable set matches Tables 4–6.

Figure 5 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 modest declines in firm size and employees relative to assets, 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 4 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 2.88 vs. 2.79), indicating that the market places a premium on knowledge intensity. Profitability deteriorates from 0.078 to –0.048 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 0.163 to 5.06, and the SG&A to sales ratio increases from 0.481 to 0.973. High θ^{know} firms also hold more liquidity, as the cash to assets ratio increases from 0.17 to 0.344, and they use less leverage, with the debt to assets ratio falling from 0.276 to 0.194. 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 left panel of Figure 6 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 5 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.

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 6 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 6 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. Profitability is modestly higher for high-θ^{org} firms, rising from 0.012 in Q1 to 0.031 in Q5, though the relationship is not strictly monotonic across the interior quintiles.

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, R&D intensity falls sharply from Q1 to Q5, while the SG&A to sales ratio rises modestly from 0.561 to 0.713, reflecting the operational, rather than R&D-driven, nature of spending among high-θ^{org} firms. Asset tangibility and leverage rise with intensity: the physical capital to assets ratio increases from 0.203 to 0.242, and the debt to assets ratio increases from 0.217 to 0.248. Liquidity moves in the opposite direction, with the cash to assets ratio falling from 0.281 to 0.212. Labor patterns also differ from those seen for knowledge or customer capital. Employees to assets increase from 0.00427 to 0.00723, while labor expense per employee declines from 104 to 88.2, indicating a larger workforce geared toward operational scale rather than specialized expertise. The right panel of Figure 6 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 3: Correlation Matrix of Intangible Capital Scores

	θ_{int}	θ_{know}	θ_{brand}	θ_{org}
θ_{int}	100.0%	62.0%	-32.2%	-33.6%
θ_{know}	62.0%	100.0%	-47.9%	-60.6%
θ_{brand}	-32.2%	-47.9%	100.0%	-34.0%
θ_{org}	-33.6%	-60.6%	-34.0%	100.0%

Notes: This table reports pairwise correlations (in percent) among the firm-year intangible intensity score θ^{int} and its three subcomponents (θ^{know}, θ^{cust}, θ^{org}), computed over the full sample of firm-year observations, 2002–2023.

To sum, these patterns across subcomponents shed light on how intangible intensity maps to firms' lifecycle dynamics and investment behavior. Table 3 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.1 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.3 reports our intensity rankings for a set of large firms in 2022. 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

Table 4: Characteristics by Knowledge Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Total Assets	6.29	6.18	6.1	6.04	5.8
Tobin's Q	2.79	3.2	3.15	3.02	2.88
Book-to-Market	0.621	0.574	0.537	0.533	0.541
Profitability	0.0783	0.0368	0.00477	-0.0224	-0.0476
Asset Tangibility	0.259	0.227	0.212	0.201	0.189
Physical Capital Investment Rate	0.269	0.323	0.349	0.367	0.347
R&D Expense / Sales	0.163	0.581	2.61	4.72	5.06
Advertising Expense / Sales	0.035	0.0363	0.0364	0.0338	0.033
SG&A Expense / Sales	0.481	0.659	0.678	0.689	0.973
Cash / Assets	0.17	0.223	0.268	0.308	0.344
Debt / Assets	0.276	0.239	0.23	0.212	0.194
Employees / Assets	0.00699	0.00532	0.00483	0.00437	0.00388
Labor Expense per Employee	78.1	90.8	93	95.8	96.2
θ^{int}	0.126	0.152	0.169	0.183	0.204
θ^{know}	0.0134	0.0343	0.055	0.077	0.113
θ^{cust}	0.0433	0.0475	0.045	0.0402	0.0359
θ^{org}	0.0682	0.0699	0.0688	0.0652	0.0546

Notes: This table presents the equal-weighted average of firm-level statistics for portfolios sorted on θ^{know}. All sorts are done within industries. All firm characteristic variables are winsorized at the 1% level. The sample period is from 2002 to 2023. The full set of firm characteristics is reported in the appendix version of this table.

placements also follow intuitive patterns with only a few exceptions.[13]

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.

5 Measuring Intangible Capital

The analysis in the preceding sections establishes that our text-based intensity scores, θ^{int} and its components, capture economically meaningful variation in firms' intangible investment behavior. However, θ measures intensity as a share: it reflects how much of a firm's disclosure is devoted to intangible-related activities, not the dollar value of the resulting capital stock. For many applications in finance and macroeconomics, researchers require estimates of intangible capital in levels. In this section, we take a first step toward combining our text-based intensity

[13]For instance, Apple falls into the lowest θ^{int} quintile within its industry (Technology and Communications), which may reflect the well-known reliance on external suppliers. It is reassuring, however, that its knowledge score ranks in the top quintile of its industry despite this.

Table 5: Characteristics by Customer Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Total Assets	5.96	6.11	6.21	6.14	5.98
Tobin's Q	3.35	3.23	3.09	2.88	2.59
Book-to-Market	0.508	0.524	0.536	0.561	0.684
Profitability	-0.0661	-0.023	0.0139	0.0389	0.0783
Asset Tangibility	0.209	0.208	0.213	0.22	0.236
Physical Capital Investment Rate	0.357	0.372	0.338	0.322	0.284
R&D Expense / Sales	7.5	5.44	2.09	0.646	0.125
Advertising Expense / Sales	0.0397	0.0297	0.0303	0.0376	0.0421
SG&A Expense / Sales	1.11	0.598	0.903	0.64	0.464
Cash / Assets	0.34	0.305	0.257	0.226	0.187
Debt / Assets	0.192	0.22	0.235	0.241	0.26
Employees / Assets	0.00495	0.00482	0.00499	0.00518	0.00518
Labor Expense per Employee	108	93.2	85.7	78.2	70.8
θ^{int}	0.187	0.18	0.166	0.156	0.145
θ^{know}	0.0826	0.0749	0.0594	0.0457	0.0306
θ^{cust}	0.0188	0.0314	0.0408	0.0518	0.0693
θ^{org}	0.0846	0.0735	0.0657	0.0586	0.0446

Notes: This table presents the equal-weighted average of firm-level statistics for portfolios sorted on θ^{cust}. All sorts are done within industries. All firm characteristic variables are winsorized at the 1% level. The sample period is from 2002 to 2023. The full set of firm characteristics is reported in the appendix version of this table.

scores with accounting data to construct improved estimates of intangible capital stocks. We begin by reviewing the measurement challenges documented in the literature, then describe our methodology, and finally present preliminary estimates of aggregate intangible capital and its components.

5.1 Challenges of Intangible Capital Measurement

The main obstacle to measuring intangible capital is the treatment of intangible investment under standard accounting rules. Unlike physical investment, which is capitalized on the balance sheet and depreciated over its useful life, measurable investment into intangibles such as R&D and advertising expenses are expensed immediately, leaving no trace on the balance sheet (Lev, 2000; Corrado et al., 2009). Moreover, because firms are not required to report these items separately, voluntary disclosure practices vary widely, leading to further measurement challenges. This accounting treatment means that standard measures of book assets systematically understate the true productive capital of intangible-intensive firms (Eisfeldt et al., 2022; Crouzet et al., 2022).

Table 6: Characteristics by Organization Capital Intensity Quintile

Variables	Q1	Q2	Q3	Q4	Q5
Log Total Assets	5.76	6.03	6.12	6.2	6.29
Tobin's Q	2.5	2.89	3.12	3.29	3.29
Book-to-Market	0.627	0.574	0.543	0.535	0.535
Profitability	0.012	0.00148	-0.00225	0.00505	0.0309
Asset Tangibility	0.203	0.201	0.217	0.223	0.242
Physical Capital Investment Rate	0.323	0.338	0.341	0.351	0.308
R&D Expense / Sales	2.38	3.59	2.4	3.04	0.975
Advertising Expense / Sales	0.0376	0.0366	0.034	0.0345	0.0299
SG&A Expense / Sales	0.561	0.661	0.678	0.799	0.713
Cash / Assets	0.281	0.29	0.274	0.259	0.212
Debt / Assets	0.217	0.225	0.229	0.233	0.248
Employees / Assets	0.00427	0.00424	0.00457	0.00515	0.00723
Labor Expense per Employee	104	94.2	86.9	81.4	88.2
θ^{int}	0.187	0.18	0.169	0.16	0.139
θ^{know}	0.0898	0.0734	0.0596	0.0464	0.0244
θ^{cust}	0.0576	0.0492	0.043	0.037	0.0256
θ^{org}	0.0388	0.0571	0.0661	0.0759	0.0886

Notes: This table presents the equal-weighted average of firm-level statistics for portfolios sorted on θ^{org}. All sorts are done within industries. All firm characteristic variables are winsorized at the 1% level. The sample period is from 2002 to 2023. The full set of firm characteristics is reported in the appendix version of this table.

The dominant approach in the literature addresses this gap by capitalizing expenditure flows using the perpetual inventory method. Eisfeldt and Papanikolaou (2013) accumulate 100% of SG&A expenses into an intangible capital stock, building on the insight of Lev and Radhakrishnan (2005) that SG&A-funded activities such as organizational processes contribute to the accumulation of organization capital. Importantly, existing approaches apply a uniform investment fraction across all firms and time periods. In reality, the share of SG&A that represents genuine investment in future productive capacity varies substantially across industries, across firms within an industry, and over time. Our text-based intensity scores provide a natural way to address this limitation: by using firm-specific, time-varying θ scores to scale SG&A into intangible investment flows before accumulation, we allow the investment fraction to reflect the heterogeneity revealed in firms' own disclosures.

5.2 Methodology

We construct intangible capital stocks using the perpetual inventory method, following the framework of Eisfeldt et al. (2022). For each firm i in year t, we define intangible investment as

the product of the firm's SG&A expenditure and its text-based intensity score. Unlike prior work, which applies a single scaling factor (typically 100% or 30%) to all firms, we use the component-level θ scores to decompose SG&A into three separate investment flows:

$$\begin{aligned} I_{i,t}^{know} &= \theta_{i,t}^{know} \times \text{SGA}_{i,t}, \\ I_{i,t}^{cust} &= \theta_{i,t}^{cust} \times \text{SGA}_{i,t}, \\ I_{i,t}^{org} &= \theta_{i,t}^{org} \times \text{SGA}_{i,t}. \end{aligned} \tag{5}$$

Each component-specific investment flow is then accumulated into a capital stock using the standard recursive formula:

$$K_{i,t}^{j} = (1-\delta)\, K_{i,t-1}^{j} + I_{i,t}^{j}, \tag{6}$$

where $j \in \{know, cust, org\}$ and $\delta = 0.20$ is the annual depreciation rate, applied uniformly to all three components following Eisfeldt et al. (2022). To initialize the recursion, we set each firm's initial capital stock at entry to

$$K_{i,0}^{j} = \frac{I_{i,1}^{j}}{g+\delta}, \tag{7}$$

where $g = 0.1$ is the approximate long-run growth rate of SG&A expenditure and $I_{i,1}^{j}$ is the first observed investment flow when the firm enters the sample. This initialization assumes the firm was in approximate steady state prior to appearing in the data (Eisfeldt et al., 2022). SG&A expenditure is obtained from Compustat (item `xsga`) and deflated by the consumer price index to express all quantities in real terms. Total intangible capital for firm i in year t is the sum of its three components:

$$K_{i,t}^{int} = K_{i,t}^{know} + K_{i,t}^{cust} + K_{i,t}^{org}. \tag{8}$$

The key distinction relative to Eisfeldt and Papanikolaou (2013) and other works is that the investment share of SG&A is no longer fixed at 100% or 30%, but instead varies across firms and over time through the θ scores. Specifically, we scale each firm's SG&A into component investment flows using its own component θ scores, with gaps between filings filled by within-firm interpolation and the most recent score carried forward.[14] This allows our capital stock estimates to reflect the firm-level heterogeneity in intangible investment intensity that we document in the preceding sections.

[14] Filings without a usable score receive the firm's interpolated or most recent score; no information from other firms enters the flows. Firm-level θ scores are highly persistent, so the filled series closely tracks the as-filed scores. See Appendix A.2 for persistence and validation analysis.

5.3 Calibrating Intensity Scores to Spending Shares

The θ scores measure the share of disclosure text devoted to each intangible category, and there is no reason for a unit of text share to correspond one-for-one to a unit of spending share. Before using the capital stocks in level-dependent applications, we therefore calibrate each component score to observed spending. Capital IQ reports disaggregated SG&A components for a subset of firms, which allows us to observe actual spending shares wherever they are disclosed. For each component $j \in \{know, cust, org\}$, we estimate the linear mapping

$$s^j_{i,t} = m^j \, \theta^j_{i,t} + u^j_{i,t}, \tag{9}$$

where $s^{know}_{i,t}$ is R&D expenditure and $s^{cust}_{i,t}$ is selling and marketing expenditure, each scaled by the sum of R&D and SG&A, and $s^{org}_{i,t}$ is the equity-pay ownership share described in Appendix A.2, a proxy rather than a direct spending share, since no accounting item isolates organization-capital expenditure. For firms without a consolidated selling-and-marketing item, customer spending falls back to the marketing item and then to the advertising item. Equation (9) is estimated at the firm-year level through the origin on the pooled sample, with Capital IQ items reported as zero treated as missing and standard errors clustered by firm. The estimated slope is the scale factor that converts a text share into a spending share.

Table 7 reports the estimates. A unit of text share maps into roughly three units of R&D spending share for knowledge capital, close to five units of selling and marketing spending for customer capital, and two units of the equity-pay outcome for organization capital, with uncentered R^2 values of 0.75, 0.70, and 0.67.[15] The calibration is deliberately parsimonious: one scale factor per component, estimated on all firms with observable spending.[16]

Figure 7 shows the corresponding validation at the industry level. The top left panel plots the industry median ratio of SG&A to sales against the industry median total score, and the relationship is tight, with an R^2 of 0.88 across the sixteen industries. The remaining panels plot industry-year median spending shares against industry-year median component θ scores. Industries that devote more disclosure text to a category spend measurably more on it wherever

[15] Estimating equation (9) separately within five broad industry groups (consumer, industrials, healthcare, technology and communications, and a residual group spanning transportation, utilities, and the remaining industries) yields similar multipliers, and the results in the paper are qualitatively unchanged when the stocks are built from the group-specific estimates. Appendix Table B.4 reports them.

[16] An alternative calibration that benchmarks the multipliers to Tobin's q rather than to observed spending yields similar results. Appendix A.5 reports the details.

Table 7: Spending-Share Calibration Multipliers

	Knowledge	Customer	Organization
m^j	2.98***	4.69***	2.04***
	(94.7)	(66.9)	(87.5)
Obs.	23,059	14,020	26,791
Firms	3,622	2,296	3,634
Uncentered R^2	0.75	0.70	0.67

Notes: This table reports the multipliers that convert component text shares into spending shares. Each column is a firm-year regression through the origin of an observed spending share on the corresponding component θ score, estimated on the pooled sample of firms with the item disclosed. Knowledge and customer spending shares are Capital IQ R&D and selling-and-marketing expenditure scaled by the sum of R&D and SG&A, with items reported as zero treated as missing, and the organization outcome is the equity-pay ownership share. All variables are winsorized at the 1st and 99th percentiles within fiscal year. t-statistics based on standard errors clustered by firm are in parentheses. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

spending is observed. The mapping is tightest for knowledge capital; the customer panel is noisier, reflecting the sparser Capital IQ coverage of selling and marketing expenditure; and the organization panel should be read with the caveat that its outcome is an equity-pay proxy rather than a direct spending share. Appendix A.2 reports the full set of validation regressions at the industry, within-industry, and within-firm level.

Throughout the remainder of the paper, we work with calibrated component stocks, $\hat{K}^j_{i,t} = m^j_{g(i)} K^j_{i,t}$, and their sum $\hat{K}^{int}_{i,t}$. Because the perpetual inventory recursion in equation (6) is linear, scaling the stock by $m^j_{g(i)}$ is equivalent to scaling the underlying investment flows, so the calibration converts the text-based stocks into dollar magnitudes comparable to spending-based measures without altering their cross-sectional or time-series variation within component and industry group.

5.4 Intangible Capital Estimates

Figure 8 plots the aggregate intangible capital share, defined as total calibrated intangible capital divided by the sum of intangible and net physical capital (Compustat `ppent`), aggregated across all firms in our sample from 2002 to 2023 (firms with missing values in a given year are treated as zero). The share rises steeply from 0.25 in 2002 to roughly 0.39 by 2009, then plateaus and drifts modestly lower through the mid-2010s before dipping to 0.38 in 2019; it rises steadily thereafter, reaching a sample peak of 0.44 by 2023. These magnitudes are consistent with the evidence in Crouzet et al. (2022) that intangible capital has become a substantial and growing

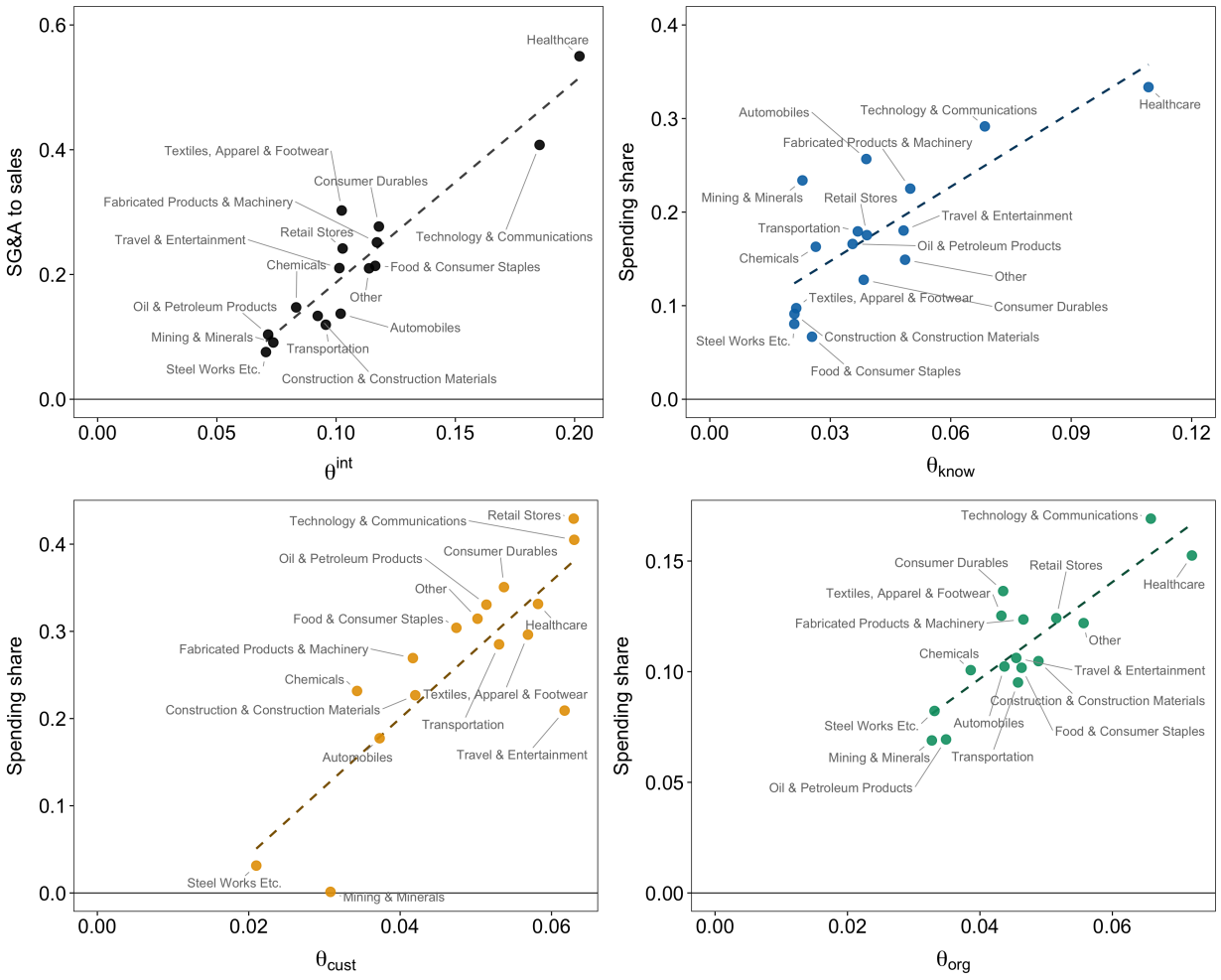


Figure 7: Validation: θ Scores and Observed Spending Shares

Notes: The top left panel plots the industry median ratio of SG&A expense to sales (Compustat) against the industry median total θ^{int} score. The remaining panels plot the industry-year median spending share against the industry-year median component θ score across industries, one panel per component. Knowledge spending is R&D and customer spending is selling and marketing expenditure from Capital IQ, each scaled by the sum of R&D and SG&A, with items reported as zero treated as missing; the organization panel uses the equity-pay ownership share as a proxy. See Appendix A.2 for the corresponding regressions.

fraction of the capital of U.S. public firms, with the most rapid accumulation occurring in the early part of the sample and renewed acceleration in the final years.

A key contribution of our approach is that the component-level θ scores allow us to measure knowledge, customer, and organization capital separately. Figure 9 plots the aggregate nominal calibrated stocks of each component over time. All three components grow steadily throughout the sample, with total intangible capital rising from approximately \$0.8 trillion in 2002 to \$4.6

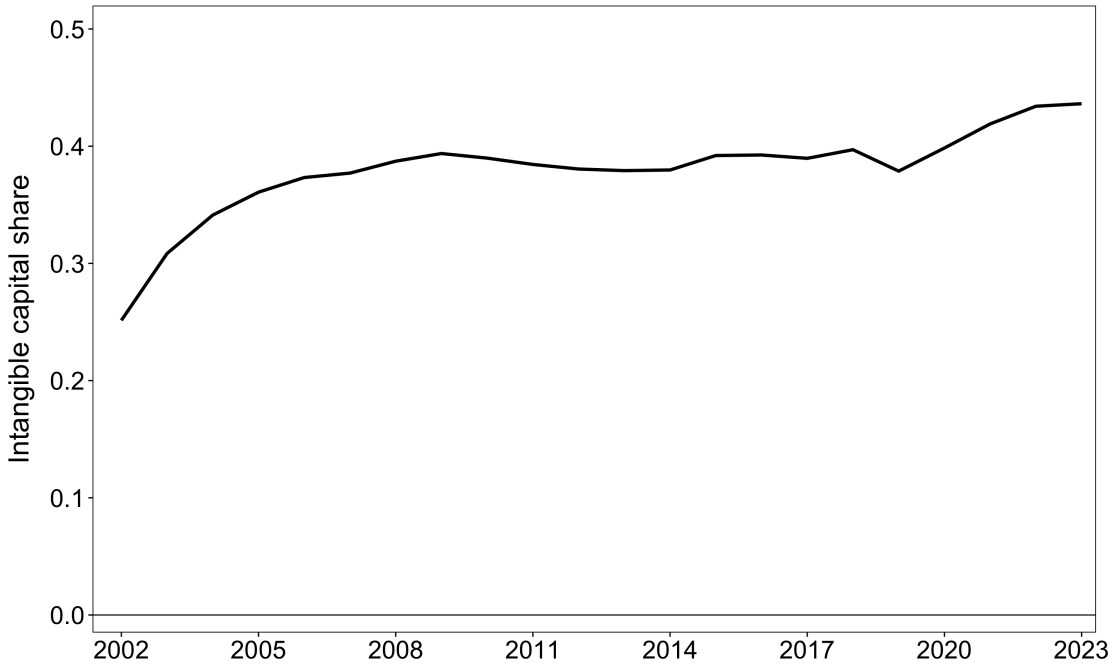


Figure 8: Aggregate Intangible Capital Share

Notes: This figure plots the ratio of total calibrated intangible capital to the sum of intangible and net physical capital (Compustat `ppent`), aggregated across all firms in the sample by fiscal year; firms with missing intangible or physical capital values in a given year are treated as zero in the aggregation rather than excluded. Intangible capital is constructed using the perpetual inventory method ($\delta = 0.20$) with component θ scores applied to SG&A expenditure and the spending-share multipliers of Section 5.3.

trillion by 2023. Customer capital is the largest component throughout the sample, at 47% of the total by 2023, reflecting both the breadth of selling and marketing activity in SG&A and the large spending-share multiplier attached to the customer score. Knowledge and organization capital are of comparable size throughout the sample, each accounting for roughly 27% of the total by 2023. Growth in the total stock accelerates markedly after 2020, consistent with the intensification of technology- and customer-acquisition-related investment among public firms in recent years.

To assess whether our θ-based capital stocks provide sharper measurement than the standard approach, Figure 10 compares the sorting discrimination of two intangible capital measures: one constructed using θ scores and one using 100% of SG&A, as in Eisfeldt and Papanikolaou (2013). For each firm characteristic, we compute the standardized quintile spread, defined as the difference in means between Q5 and Q1 divided by the pooled standard deviation. The θ-based measure generates notably larger spreads for variables most closely tied to the composition of intangible investment, including knowledge spending share and brand spending share. It also produces

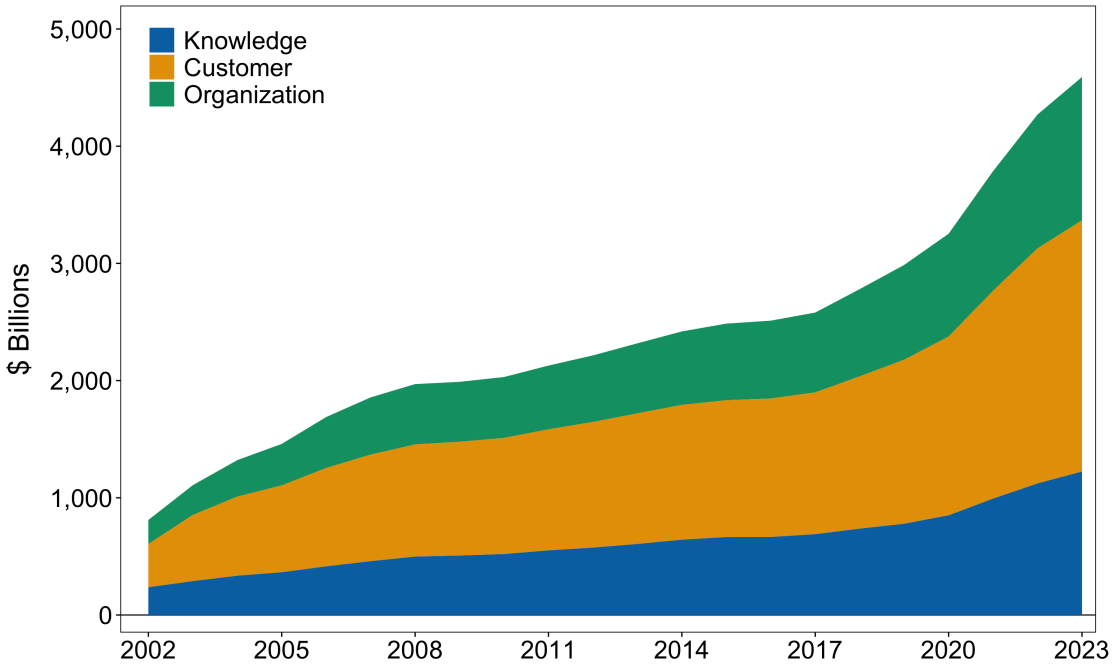


Figure 9: Aggregate Intangible Capital by Component

Notes: This figure plots aggregate nominal calibrated intangible capital stocks by component (knowledge, customer, and organization capital) for all firms in the sample. Each component is accumulated using the perpetual inventory method ($\delta = 0.20$) with component-specific θ scores applied to SG&A expenditure and scaled by the spending-share multipliers of Section 5.3.

modestly larger spreads for cash-to-assets, debt-to-assets, and SG&A-to-sales. By contrast, the 100% SG&A method produces a larger spread for asset tangibility, which is expected: capitalizing all of SG&A mechanically inflates the intangible stock for firms with high operating expenses, creating stronger separation along the tangible-intangible margin regardless of whether those expenses represent genuine investment. Overall, the comparison suggests that incorporating text-based intensity scores into the perpetual inventory method improves the ability to distinguish firms along the dimensions most directly related to intangible capital composition.

Our perpetual inventory method already applies firm-specific θ scores, filled by within-firm interpolation, rather than industry-level averages. This firm-level variation draws further signal from within-industry heterogeneity and sharpens the measurement of intangible capital for applications in asset pricing and corporate finance.

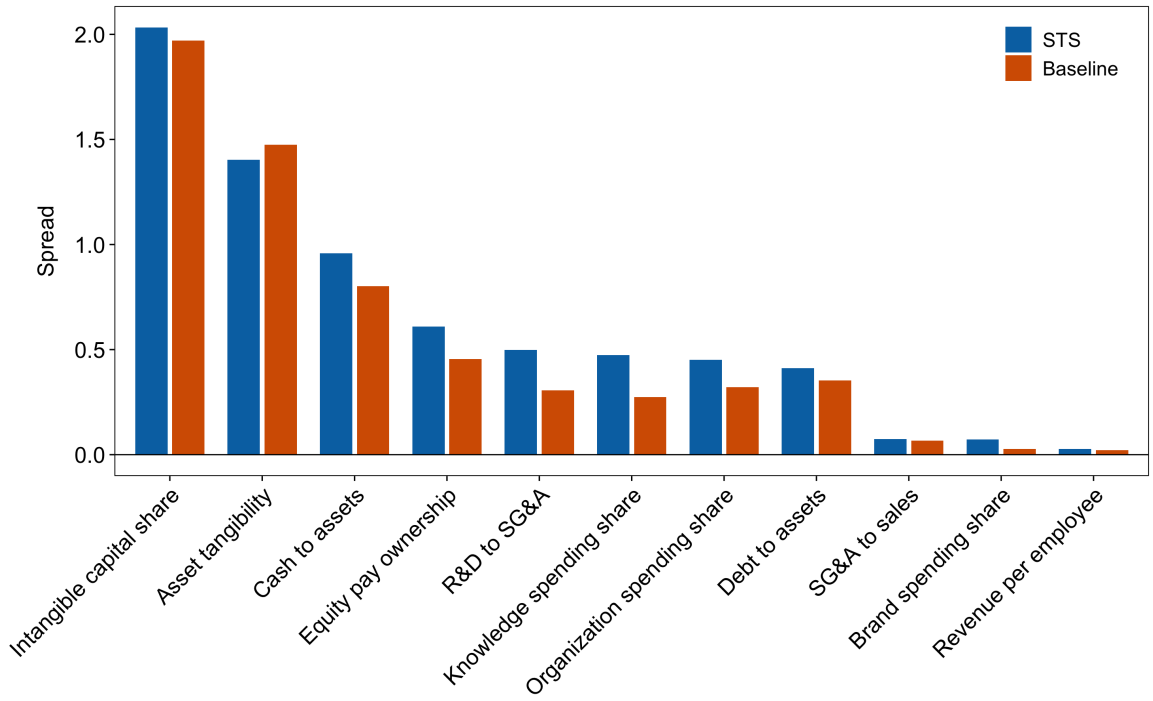


Figure 10: Sorting Discrimination: θ-Based vs. 100% SG&A Capital Stocks

Notes: This figure plots the standardized quintile spread, $(Q5 - Q1)/\sigma$, for firms sorted on the ratio of intangible to total capital, comparing intangible capital constructed with θ scores (STS, blue bars) against a baseline that capitalizes 100% of SG&A following Eisfeldt and Papanikolaou (2013) and Eisfeldt et al. (2022) (Baseline, red bars). All sorts are within industries.

6 Intangible Capital and the Valuation Wedge

If book values measured all of a firm's capital and market values contained no rents, the ratio of the market value of equity to its book value would be close to one. This identity does not hold in the data. The average market-to-book ratio in our sample is 3.9 and varies widely across firms. Two economic forces drive this wedge. First, internally generated intangible capital is expensed rather than capitalized, so the book value of equity omits a growing share of firms' productive capital (Eisfeldt et al., 2022; Crouzet et al., 2022). Second, market values embed franchise value: the present value of rents and growth opportunities that no measure of installed capital, however complete, should be expected to capture (Miller and Modigliani, 1961). We ask how much of the cross-sectional variation in valuation ratios these two forces explain, using the calibrated component stocks together with a simple franchise-value proxy. The exercise is an accounting decomposition of the valuation wedge rather than a structural test: once rents and growth options enter, theory does not predict unit or additive coefficients on measured capital (Crouzet and

Eberly, 2023), so our diagnostic is explanatory power rather than the value of any particular coefficient.

We construct franchise value from a growing-perpetuity valuation model,

$$FV_{i,t} = \frac{\max\left(E_{i,t},\, 0\right)}{r_t - g_{i,t}}, \tag{10}$$

where $E_{i,t}$ is earnings (income before extraordinary items), r_t is the cost of equity under a unit market beta, set to the ten-year Treasury yield plus a 6% equity premium, and $g_{i,t}$ is the firm's trailing three-year sales growth, with industry-year medians substituted for firms with short histories.[17]

We then estimate nested specifications of the form

$$\frac{ME_{i,t}}{BE_{i,t}} = \alpha + \sum_{j \in \{know,\, cust,\, org\}} \beta_j \frac{\hat{K}^j_{i,t}}{BE_{i,t}} + \lambda \frac{FV_{i,t}}{BE_{i,t}} + \tau_t + \varepsilon_{i,t}, \tag{11}$$

where $ME_{i,t}$ and $BE_{i,t}$ are the market and book value of equity, τ_t denotes year fixed effects, and standard errors are clustered by firm. The sample contains firms with positive book equity, assets, and sales from 2003 to 2023, and all ratios are winsorized at the 1% level. We omit firm fixed effects because the object of interest is the cross-firm dispersion in valuation levels. We also replace the three components with the total calibrated stock $\hat{K}^{int}_{i,t}$, or omit the intangible terms, or omit franchise value; the diagnostic across columns is the within-year R^2.

Table 8 reports the results. Franchise value on its own explains 0.12 of the within-year variation in market-to-book (column 1). The total intangible stock alone explains 0.27, and adding franchise value to it raises the explanatory power to 0.38 (columns 2–3): omitted intangible capital alone accounts for over a quarter of the cross-sectional variation. Decomposing the stock into the three components raises the within R^2 to 0.29 on its own and to 0.40 with franchise value (columns 4–5). The two components of the wedge are largely complementary rather than substitutes: intangible capital contributes most of the explained variation, and franchise value adds roughly eleven percentage points of R^2 on top of it.

[17] Equation (10) floors franchise value at zero for the roughly 38% of firm-years with negative earnings, so the estimates are a conservative lower bound on the contribution of franchise value. We estimate growth from sales rather than earnings because, for the firms where the two diverge most, trailing earnings growth is undefined or floored out of the formula entirely. Growth rates are winsorized within year and capped so that $r_t - g_{i,t}$ stays above 2%. A constant $r - g$ of 0.06, an earnings-based growth rate, and a sales-based proxy that is strictly positive for loss-making firms all yield the same component pattern.

Table 8: Intangible Capital and the Cross-Section of Market-to-Book

	Market-to-book				
	(1)	(2)	(3)	(4)	(5)
Intangible capital $/BE$		0.86***	0.84***		
		(29.9)	(32.1)		
Knowledge $/BE$				0.88***	1.20***
				(5.9)	(8.9)
Customer $/BE$				0.25**	0.18**
				(2.4)	(2.0)
Organization $/BE$				2.17***	1.90***
				(9.5)	(10.1)
Franchise value $/BE$	0.34***		0.33***		0.33***
	(19.3)		(24.0)		(24.8)
Year FE	✓	✓	✓	✓	✓
Obs.	45,003	45,003	45,003	45,003	45,003
Within R^2	0.118	0.272	0.382	0.288	0.398

Notes: This table reports estimates of equation (11). The dependent variable is the ratio of the market value to the book value of equity. Intangible capital stocks are the calibrated text-based stocks of Section 5.3; franchise value is constructed from equation (10). The sample contains Compustat non-financial firms with positive book equity, assets, and sales, 2003–2023; all ratios are winsorized at the 1% level. All columns are estimated on the common sample of firm-years with nonmissing calibrated stocks; the franchise-value ratio alone is available for 57,923 firm-years. t-statistics based on standard errors clustered by firm are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

The decomposition also reveals that knowledge and organization capital enter with large and highly significant coefficients in the full specification (column 5), while the customer component is small and statistically insignificant. The decomposition across components should be interpreted with caution because customer and organization capital are close substitutes, making their individual contributions sensitive to the order in which they enter the specification. The qualitative result is unchanged across orderings: knowledge and organization capital account for essentially all of the incremental explanatory power, while customer capital – the margin emphasized by He et al. (2024) at the industry level – adds little once the other two are included. For firm-level valuation, the economically relevant margin is therefore knowledge versus organization capital, a pattern that can only be uncovered with firm-level decomposition.

We next study how the decomposition fares against the standard determinants of valuation ratios (Rhodes-Kropf et al., 2005). Table 9 augments the full specification with the firm characteristics a conventional valuation regression would include: size, profitability, leverage, asset turnover, age, and dividend status, and then progressively absorbs industry and industry-by-year

Table 9: The Valuation Wedge and Standard Determinants

	Market-to-book			
	(1)	(2)	(3)	(4)
Knowledge $/BE$	1.20***	1.02***	0.93***	0.96***
	(8.9)	(7.5)	(6.8)	(7.0)
Customer $/BE$	0.18**	0.32***	0.32***	0.31***
	(2.0)	(3.5)	(3.4)	(3.4)
Organization $/BE$	1.90***	1.84***	1.88***	1.86***
	(10.1)	(10.0)	(10.2)	(10.2)
Franchise value $/BE$	0.33***	0.36***	0.36***	0.36***
	(24.8)	(24.1)	(24.0)	(24.0)
Controls		✓	✓	✓
Year FE	✓	✓	✓	
Industry FE			✓	
Year × Industry FE				✓
Obs.	45,003	42,761	42,761	42,761
Within R^2	0.398	0.436	0.427	0.430

Notes: This table reports estimates of equation (11) augmented with standard valuation determinants. Column (1) repeats the full specification of Table 8. Columns (2)–(4) add the controls of standard cross-sectional valuation regressions (Rhodes-Kropf et al., 2005): the logarithm of book assets, return on assets, book leverage, sales over assets, the logarithm of one plus firm age, and a dividend-payer indicator; coefficient estimates on the controls are suppressed. All ratios are winsorized at the 1% level. t-statistics based on standard errors clustered by firm are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

fixed effects. Three results stand out. First, the intangible components and franchise value are remarkably stable: in column 4, where identification comes exclusively from cross-firm differences in accumulated investment within an industry and year, the coefficients on knowledge, customer, and organization capital and on franchise value are essentially unchanged from the baseline, and all significant at the 1% level. All three components and franchise value enter positively in every specification. Second, the standard determinants add little explanatory power: the within-R^2 rises from 0.398 to 0.436, so conventional characteristics contribute roughly a tenth of what the intangibles and franchise value contribute on their own. Third, the customer component, the weakest of the three in the baseline, strengthens once we control for firm characteristics. The reason is suppression: customer-intensive firms trade at lower multiples because they are high-turnover and profitable.[18]

If the omitted-capital component of the wedge is economically important, adding the calibrated

[18]For profitable firms, franchise value combines the return on equity with the firm's estimated growth rate, so the information it adds relative to a simple profitability control comes from the growth capitalization rather than from profitability per se.

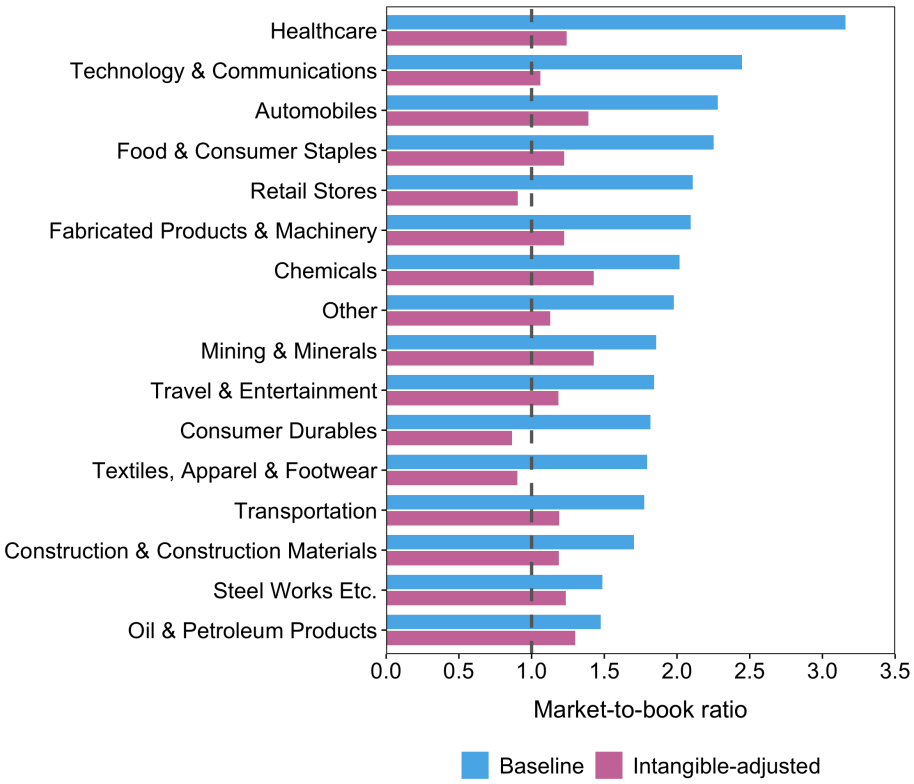


Figure 11: Industry Market-to-Book Before and After the Intangible Adjustment

Notes: This figure plots, for each industry, the time average (2002–2023) of the industry-year median market-to-book ratio computed on book equity as reported (Baseline, the market-to-book ratio from accounting variables only) and on book equity plus the calibrated total intangible stock $\hat{K}_{i,t}^{int}$ (Intangible-adjusted). The dashed line marks a market-to-book ratio of one. Utilities and financial firms are excluded from the sample.

intangible stock back to book equity should move valuation ratios toward one. Figure 11 performs this accounting exercise industry by industry, computing the median market-to-book ratio in each year before and after adding $\hat{K}_{i,t}^{int}$ to book equity and averaging over time. The adjustment moves the typical industry from a median market-to-book ratio of 1.9 to 1.2, and only three of the sixteen industries fall below one. The discipline in this exercise is that while adding any large stock to the denominator mechanically compresses the ratio, the calibrated stocks move most industries close to one without overshooting it. This contrasts with capitalizing 100% of SG&A, which lowers the market-to-book ratio to 0.77 and pushes 88% of industries below one, overstating omitted capital. The calibrated text-based stocks thus remove most of the wedge while leaving room for franchise value to account for the remaining gap between market and replacement values.

7 Intangible Capital and Capital Structure

Financing is another corporate outcome for which the composition and quantity of intangible capital should matter. Collateral-based theories predict that firms whose capital is intangible borrow less, because intangible assets pledge poorly (Rampini and Viswanathan, 2013), while the modern lending literature documents that most large-firm borrowing is in fact cash flow-based rather than asset-based (Lian and Ma, 2021; Kermani and Ma, 2023), which weakens the mechanical link between tangibility and debt capacity. Quantitatively, the secular rise of intangible capital has been linked to shrinking debt capacity and to the rise in corporate cash holdings (Falato et al., 2022), and intangible capital embodied in a firm's employees is financed through deferred compensation that crowds out debt (Sun and Xiaolan, 2019). The sharpest evidence in our sample is at the extensive margin. The share of firms with no debt at all rises monotonically from 4% in the lowest intangible-intensity quintile to 32% in the highest, the zero-leverage pattern documented by Strebulaev and Yang (2013). We next turn to the intensive margin and put the calibrated stocks directly into a standard capital-structure regression. We estimate

$$y_{i,t} = \beta \frac{\hat{K}^{int}_{i,t-1}}{A_{i,t-1}} + \boldsymbol{\gamma}' \mathbf{X}_{i,t-1} + \mu_{j(i)\times t} + \varepsilon_{i,t}, \tag{12}$$

where $y_{i,t}$ is book leverage or the cash-to-assets ratio, $\mathbf{X}$ contains the standard determinants (Rajan and Zingales, 1995; Frank and Goyal, 2009; Titman and Wessels, 1988), $\mu_{j\times t}$ denotes industry-by-year fixed effects, and standard errors are clustered by firm. We restrict the sample to firm-years with positive book equity, as in Fama and French (1992).

Table 10 reports the leverage results. The classic determinants behave largely as in the literature: tangibility and size raise book leverage, while profitability, firm age, and dividend-paying status lower it (column 1). Adding the intangible stock yields a negative association that is strongest for market leverage (column 5): a one-standard-deviation increase in $\hat{K}^{int}/A$ is associated with market leverage lower by about 2.7 percentage points. The book-leverage association (column 2) is of similar size, at about 2.3 percentage points per standard deviation. Intangible capital adds to rather than displaces the classical channels, as the tangibility coefficient remains the dominant determinant. Two additional patterns help interpret the estimates. First, the relation is a between-firm phenomenon. With firm fixed effects, the coefficient turns positive, indicating that firms that accumulate intangible capital over time do not visibly delever (column 4). The cross-sectional relation therefore describes which firms have low leverage, rather than how leverage changes as a given firm invests. Second, the component decomposition (column 3) attributes the negative

Table 10: Intangible Capital and Leverage

	(1)	(2)	(3)	(4)	(5)
$\hat{K}^{int}/A$		−0.027***		0.011***	−0.031***
		(−9.6)		(3.2)	(−10.5)
Knowledge/A			−0.003		
			(−0.4)		
Customer/A			0.001		
			(0.2)		
Organization/A			−0.057***		
			(−4.2)		
Tangibility	0.246***	0.212***	0.211***	0.213***	0.247***
	(18.2)	(13.9)	(14.5)	(10.1)	(14.3)
log(Assets)	0.035***	0.032***	0.031***	0.052***	0.026***
	(28.7)	(22.2)	(23.0)	(14.3)	(16.0)
Market-to-book	0.005***	0.005***		0.004***	−0.004***
	(16.1)	(15.4)		(14.4)	(−12.6)
Profitability	−0.049***	−0.085***		−0.074***	−0.116***
	(−6.8)	(−9.2)		(−7.0)	(−13.1)
Firm age	−0.012***	−0.013***	−0.014***	0.017***	−0.007**
	(−5.1)	(−5.0)	(−7.0)	(2.9)	(−2.4)
Dividend payer	−0.036***	−0.029***	−0.031***	−0.008*	−0.051***
	(−7.4)	(−5.4)	(−6.1)	(−1.8)	(−8.8)
Dependent variable	Book	Book	Book	Book	Market
Firm FE				✓	
Year FE				✓	
Industry × Year FE	✓	✓	✓		✓
Obs.	49,667	38,644	43,244	38,644	38,644
Within R^2	0.171	0.176	0.143	0.071	0.139

Notes: This table reports estimates of equation (12). The dependent variable is book leverage, $(dltt + dlc)/at$, except in column (5), where it is market leverage, $(dltt + dlc)/(dltt + dlc + ME)$. Intangible stocks are scaled by assets. Column (3) omits market-to-book and profitability, which are related to the intangible stocks by construction. All right-hand-side variables are lagged one year; firm age enters as the logarithm of one plus age; continuous ratios are winsorized at the 1% level within year. The sample contains non-financial Compustat firms with positive book equity, 2003–2023. t-statistics based on standard errors clustered by firm are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

association to the organization component, despite the fact that leverage rises with organization intensity in the unconditional sorts of Table 5. This implies that organization-intensive firms are larger and more tangible, and both characteristics independently raise leverage. Once these characteristics are held fixed, the residual association is negative.

Table 11 turns to cash holdings, where the composition of intangible capital matters most sharply. In the standard specification of Opler et al. (1999), the intangible stock enters positively

and strongly (column 2): a one-standard-deviation increase in $\hat{K}^{int}/A$ is associated with cash holdings that are 4.3 percentage points higher, about 19% of a standard deviation, consistent with the precautionary demand of firms whose capital cannot be pledged (Bates et al., 2009; Falato et al., 2022). Column 3 shows that the association is dominated by knowledge capital, with organization capital a distant second and customer capital small and negative. This knowledge-first ordering mirrors the valuation results and is consistent with knowledge capital being the least collateralizable and most information-sensitive component. As with leverage, the level relation does not survive firm fixed effects (column 4), so it describes persistent differences across firm types rather than within-firm dynamics. The same loss-making, knowledge-intensive firms for which replacement value provides the weakest valuation anchor are also the firms holding the most cash, consistent with costly external finance both making growth value difficult to price and inducing firms to self-insure with liquidity. A sharper test of the collateral mechanism would split debt into asset-based and cash-flow-based claims, which is beyond the scope of this paper (Lian and Ma, 2021).

Table 11: Intangible Capital and Cash Holdings

	(1)	(2)	(3)	(4)
$\hat{K}^{int}/A$		0.049***		−0.004
		(14.0)		(−1.0)
Knowledge/A			0.135***	
			(12.4)	
Customer/A			−0.014*	
			(−1.8)	
Organization/A			0.097***	
			(5.7)	
log(Assets)	−0.031***	−0.024***	−0.023***	−0.050***
	(−24.0)	(−16.3)	(−15.7)	(−16.2)
Market-to-book	0.003***	0.003***		0.000**
	(13.6)	(10.3)		(2.1)
Cash flow/A	−0.162***	−0.060***	−0.000	0.054***
	(−16.5)	(−4.7)	(−0.0)	(5.2)
NWC/A	−0.442***	−0.399***	−0.378***	−0.262***
	(−36.7)	(−31.4)	(−29.4)	(−15.3)
Capex/A	−0.750***	−0.598***	−0.574***	−0.240***
	(−23.5)	(−17.4)	(−17.0)	(−10.0)
Dividend payer	−0.017***	−0.014***	−0.016***	−0.001
	(−3.8)	(−3.1)	(−3.5)	(−0.4)
Firm FE				✓
Year FE				✓
Industry × Year FE	✓	✓	✓	
Obs.	55,812	42,736	42,736	42,736
Within R^2	0.286	0.268	0.275	0.069

Notes: This table reports estimates of equation (12) with the cash-to-assets ratio as the dependent variable and the control set of Opler et al. (1999). Cash flow is operating cash flow over assets; NWC is net working capital excluding cash. Column (3) omits market-to-book, which is related to the intangible stocks by construction. All right-hand-side variables are lagged one year; continuous ratios are winsorized at the 1% level within year. The sample contains non-financial Compustat firms with positive book equity, 2003–2023. t-statistics based on standard errors clustered by firm are in parentheses. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

8 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.

The applications in Sections 6 and 7 illustrate the payoff to measuring composition: the calibrated component stocks, combined with a simple franchise-value proxy, explain a large share of the cross-section of market-to-book ratios, and intangible capital is a robust correlate of financing choices, most sharply the precautionary cash holdings of knowledge-intensive firms. These measures open further avenues for research, including whether intangible intensity predicts future operating performance and investment, and how the composition of intangible capital interacts with debt structure and the pricing of 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

Aboody, David and Baruch Lev, "The Value Relevance of Intangibles: The Case of Software Capitalization," *Journal of Accounting Research*, 1998, *36*, 161–191.

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.

Bates, Thomas W., Kathleen M. Kahle, and René M. Stulz, "Why Do U.S. Firms Hold So Much More Cash than They Used To?," *The Journal of Finance*, 2009, *64* (5), 1985–2021.

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.

Chan, Louis K. C., Josef Lakonishok, and Theodore Sougiannis, "The Stock Market Valuation of Research and Development Expenditures," *The Journal of Finance*, 2001, *56* (6), 2431–2456.

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, "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.

Crouzet, Nicolas and Janice C. Eberly, "Rents and Intangible Capital: A Q+ Framework," *The Journal of Finance*, 2023, *78* (4), 1873–1916.

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

Daniel, Kent and Sheridan Titman, "Market Reactions to Tangible and Intangible Information," *The Journal of Finance*, 2006, *61* (4), 1605–1643.

David, Alexander, Amir Hosseini, and Anup Srivastava, "Is Intangibles Talk Informative About Future Returns? Evidence from 10-K Filings," Working Paper, University of Calgary 2023.

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*, 2025, *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.

Fama, Eugene F. and Kenneth R. French, "The Cross-Section of Expected Stock Returns," *The Journal of Finance*, 1992, *47* (2), 427–465.

Filipovic, Zoran M. and Alexander F. Wagner, "The Intangibles Song in Takeover Announcements: Good Tempo, Hollow Tune," *The Review of Financial Studies*, 2026. Forthcoming.

Frank, Murray Z. and Vidhan K. Goyal, "Capital Structure Decisions: Which Factors Are Reliably Important?," *Financial Management*, 2009, *38* (1), 1–37.

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*, 2025, *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.

Kermani, Amir and Yueran Ma, "Asset Specificity of Nonfinancial Firms," *The Quarterly Journal of Economics*, 2023, *138* (1), 205–264.

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

_ **and Paul Zarowin**, "The Boundaries of Financial Reporting and How to Extend Them," *Journal of Accounting Research*, 1999, *37* (2), 353–385.

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

_ **and Theodore Sougiannis**, "The Capitalization, Amortization, and Value-Relevance of R&D," *Journal of Accounting and Economics*, 1996, *21* (1), 107–138.

Lian, Chen and Yueran Ma, "Anatomy of Corporate Borrowing Constraints," *The Quarterly Journal of Economics*, 2021, *136* (1), 229–291.

Loughran, Tim and Bill McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, 2011, *66* (1), 35–65.

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.

Miller, Merton H. and Franco Modigliani, "Dividend Policy, Growth, and the Valuation of Shares," *The Journal of Business*, 1961, *34* (4), 411–433.

Opler, Tim, Lee Pinkowitz, René M. Stulz, and Rohan Williamson, "The Determinants and Implications of Corporate Cash Holdings," *Journal of Financial Economics*, 1999, *52* (1), 3–46.

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.

Rajan, Raghuram G. and Luigi Zingales, "What Do We Know about Capital Structure? Some Evidence from International Data," *The Journal of Finance*, 1995, *50* (5), 1421–1460.

Rampini, Adriano A. and S. Viswanathan, "Collateral and Capital Structure," *Journal of Financial Economics*, 2013, *109* (2), 466–492.

Rhodes-Kropf, Matthew, David T. Robinson, and S. Viswanathan, "Valuation Waves and Merger Activity: The Empirical Evidence," *Journal of Financial Economics*, 2005, *77* (3), 561–603.

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

Strebulaev, Ilya A. and Baozhong Yang, "The Mystery of Zero-Leverage Firms," *Journal of Financial Economics*, 2013, *109* (1), 1–23.

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

Titman, Sheridan and Roberto Wessels, "The Determinants of Capital Structure Choice," *The Journal of Finance*, 1988, *43* (1), 1–19.

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

Appendix

"Intangible Intensity"

July 2026

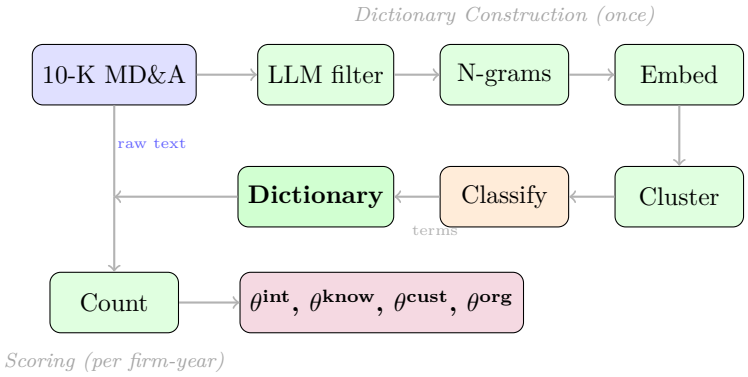
A Additional Details on Methodology **2**
A.1 Semantic Theme Scoring (STS) Methodology 2
A.2 Validation of Industry-Level θ Scores 7
A.3 Validation Against Direct LLM Scoring 9
A.4 LLM Scoring Prompt for STS Validation 11
A.5 Calibration Using Tobin's q 13

B Additional Figures and Tables **14**

A Additional Details on Methodology

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, with canonical NLP processing framework for n-gram corpus generation.

2. Compute corpus-wide frequencies and keep the top N n-grams (in the slides, $N = 30{,}000$) 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. Cutting-edge transformer embedding models are less prone to anisotropy problem. Yet, residual concentration of variance could potentially remain along a few directions. Whitening corrects this by equalizing the variance across principal components, so downstream distance-based clustering becomes well-conditioned. 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 ($K = 1000$). This creates many fine-grained groups of semantically similar n-grams.

2. Construct a graph whose nodes are the micro-clusters; connect nearby micro-clusters with an edge whose weight aggregates how often their member n-grams appear in each other's nearest-neighbor lists.

3. Apply a community-detection algorithm (Leiden, with the Constant Potts Model objective) to this graph to merge micro-clusters into a few hundred "communities" of n-grams.

Thirty 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, organization capital and non-intangible) 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**
- **Unknown**

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*, *non-intangible-investment*, and *unknown*.

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.

The LLM-community classification prompt is:

```
You are a financial analyst with a PhD in financial economics, and a CPA.

You will be given a set of n-grams (bigrams and trigrams) from firms' annual
    reports (item 7) that were grouped together using topic modeling.  The set of n
    -grams represents one topic.

{Set of N-grams}

Your task is to assign the entire topic to one of the following top-level
    categories:

        •       Intangible investment -- a cost expected to build or grow one of
            the three categories of intangible assets, namely knowledge capital,
            brand or customer capital, and organization capital.  Examples of
            expenses that grow knowledge capital are R&D, IT, technology, data, and
             software expenses, and labor expenses associated with R&D or research
            and development.  Examples of expenses that grow brand or customer
            capital are advertising, marketing, and selling expenses as well as
            distribution costs and customer acquisition costs (CAC).  Examples of
            expenses that grow organization capital include consulting fees,
            compensation of high-skilled employees, equity compensation, key
```

```
            personnel expenses , management personnel, training expenses,
            collaborative arrangements, and travel expenses.  These are all
            examples of costs expected to grow intangible assets.
        •       Not intangible investment -- a cost not expected to contribute to
            intangible assets.  Examples of expenses that are not expected to grow
            any intangible assets are depreciation, amortization, litigation
            expenses, bad debt expenses, transactions costs, overhead,
            restructuring costs due to cost cutting or otherwise, rental expenses,
            and effects of foreign exchange rates or inflation.
        •       Unknown -- if the cluster does not contain enough information to
            determine the category

If you classify the cluster as Intangible investment, then also assign it to one
    of the 3 subcategories, knowledge capital, brand or customer capital, or
    organization capital.  Recall:

        •       Knowledge capital refers to costs intended to create or improve
            reusable technical knowledge, software, data assets, engineering
            capability, product-development capability, patents, algorithms,
            internal tools, or technology platforms that can support future
            innovation or operations.
        •       Brand or customer capital refers to costs intended to build brand
            awareness, demand generation, customer relationships, customer lists,
            customer acquisition, customer retention, channel reach, distribution
            capability, or sales enablement that can benefit future periods.
        •       Organization capital refers to costs intended to improve
            managerial capability, workforce capability, coordination, operating
            routines, internal processes, implementation capability, organizational
             design, change management, collaboration, or firm-specific know-how
            embedded in the organization. Organization capital may include training
            , employee development, high-skill labor, recruiting specialized talent
            , management capability development, process redesign, implementation
            consulting, change-management efforts, and professional services that
            clearly support system, process, or organizational capability building.

Output your answer in this JSON format:
{
  "category": "<Intangible investment | Not intangible investment | Unknown>",
  "subcategory": "<Knowledge capital | Brand or customer capital | Organization
      capital | null>"
}
```

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.

A.2 Validation of Industry-Level θ Scores

This appendix examines whether industry-level θ scores predict actual intangible spending shares, using Capital IQ data on the composition of SG&A expenses. For each component $k \in \{know, cust, org\}$, we define the spending share as the ratio of component-specific expenditure to total SG&A. For organization capital expenditure share, due to lack of accounting standards in reporting, there is no separate item that can be used. Therefore we use equity pay variables, in particular a share of ownership in equity pay, as a proxy for organization capital expenditure share. We estimate three panel specifications that progressively absorb more variation.

Calibration inputs. The spending shares used in the calibration of Section 5.3 are constructed from Capital IQ as follows. Knowledge spending is the R&D expenditure item. Customer spending is the consolidated selling-and-marketing item where reported; for firms without it, the marketing item; and for firms without either, the advertising item. Each spending item is treated as missing when reported as zero or negative, and the denominator is the sum of R&D and SG&A. The organization outcome is the equity-pay ownership share expressed as a fraction, with zeros retained, and component θ scores enter with negative values treated as missing. All variables are winsorized at the 1st and 99th percentiles within fiscal year, and equation (9) is estimated through the origin with standard errors clustered by firm. The validation regressions in this appendix measure customer spending with the consolidated selling-and-marketing item only; the calibration additionally uses the marketing and advertising fallbacks to maximize coverage.

Table A.1 reports the inter-industry results. We estimate

$$\bar{\text{Share}}^{k}_{j,t} = \tau_t + \lambda\, \bar{\theta}^{k}_{j,t} + u_{j,t},$$

where $\bar{\text{Share}}^{k}_{j,t}$ and $\bar{\theta}^{k}_{j,t}$ are the industry-year means for industry j in year t, and τ_t denotes time fixed effects. Capital IQ spending items reported as zero are treated as missing throughout, matching the convention used for the calibration in Section 5.3. Industry-average θ scores are strong predictors of industry-average spending shares, with R^2 values of 0.49 for knowledge, 0.34 for brand, and 0.64 for the organization proxy. Figure 7 in the main text illustrates these relationships visually.

Table A.1: Inter-Industry Variation (Industry-Year Means)

	Knowledge Share	Brand Share	Equity Pay
θ^k	2.233***	2.092**	2.120***
	(7.81)	(2.06)	(5.76)
R^2	0.491	0.342	0.644
Obs.	344	289	288

Notes: Spending shares are constructed with the zeros-as-missing (`no_zero=True`) convention: each raw CIQ spending item (R&D, SG&A, Sales & Marketing) is set to missing when it equals exactly zero before forming the ratio. Standard errors are double-clustered at the industry and year level. t-statistics in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Tables A.2 and A.3 decompose this predictive power further. Table A.2 estimates a within-industry specification:

$$\text{Share}^{k}_{i,t} = \mu_j + \tau_t + \beta\, \theta^{k}_{i,t} + \epsilon_{i,t},$$

where μ_j and τ_t denote industry and year fixed effects. The within R^2 values of 0.28 for knowledge, 0.09 for brand, and 0.03 for the organization proxy indicate that knowledge-capital scores in particular explain a substantial share of firm-level spending variation even after absorbing industry and year effects. Table A.3 estimates a within-firm specification:

$$\text{Share}^{k}_{i,t} = \delta_i + \tau_t + \phi\, \theta^{k}_{i,t} + \varepsilon_{i,t},$$

where δ_i denotes firm fixed effects. The within R^2 values drop to 0.04, 0.01, and effectively zero, confirming that θ scores are highly persistent within firms and capture primarily cross-sectional rather than time-series variation. This persistence motivates our use of firm-level θ scores, filled by within-firm interpolation, in the perpetual inventory method: because the scores vary little within a firm over time, the filled series closely tracks the as-filed scores while preserving the firm-level heterogeneity that industry averages would discard.

Table A.2: Within-Industry Variation (Industry & Time FE)

	Knowledge Share	Brand Share	Equity Pay
θ^k	1.801***	1.896***	0.537***
	(36.64)	(13.98)	(13.21)
Within R^2	0.285	0.087	0.033
Obs.	23,060	12,079	26,791

Notes: Spending shares are constructed with the zeros-as-missing (`no_zero=True`) convention: each raw CIQ spending item (R&D, SG&A, Sales & Marketing) is set to missing when it equals exactly zero before forming the ratio. Firm-year panel with industry and year fixed effects. Standard errors are clustered by firm. t-statistics in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Table A.3: Within-Firm Variation (Firm & Time FE)

	Knowledge Share	Brand Share	Equity Pay
θ^k	0.609***	0.373***	0.027
	(9.28)	(4.67)	(0.84)
Within R^2	0.038	0.009	0.000
Obs.	23,060	12,079	26,791

Notes: Spending shares are constructed with the zeros-as-missing (`no_zero=True`) convention: each raw CIQ spending item (R&D, SG&A, Sales & Marketing) is set to missing when it equals exactly zero before forming the ratio. Firm-year panel with firm and year fixed effects. Standard errors are clustered by firm. t-statistics in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

A.3 Validation Against Direct LLM Scoring

This appendix compares STS against an alternative in which a frontier LLM is prompted directly to classify each firm's intangible investment intensity. Under this "guided oracle" approach, a model such as GPT, Claude Sonnet, Claude Haiku, or Gemini reads the full Item 7 text and assigns a categorical rating of low, medium, or high for each of the three intangible capital components. This approach is appealing because it leverages the full reasoning capacity of the model, but it raises concerns about reproducibility, cost, and scalability that we examine here.

We conduct a benchmarking exercise using 50 10-K Item 7 documents drawn from the 2023 fiscal year. These documents are selected as the largest firms by market capitalization within each industry. Each document is scored by four models: GPT, Claude Sonnet, Claude Haiku, and Gemini. The scoring prompt instructs the model to read the full Item 7 text, identify all passages discussing SG&A or operating expenses, and assign a level of low, medium, or high for knowledge, customer, and organization capital based on the depth and prominence of the discussion.[1] To assess within-model reliability, we run Haiku twice independently at the same temperature setting.

Figure A.1 plots STS θ scores against the LLM-assigned levels for each model and component.

[1]The full scoring prompt is reproduced in Appendix A.4.

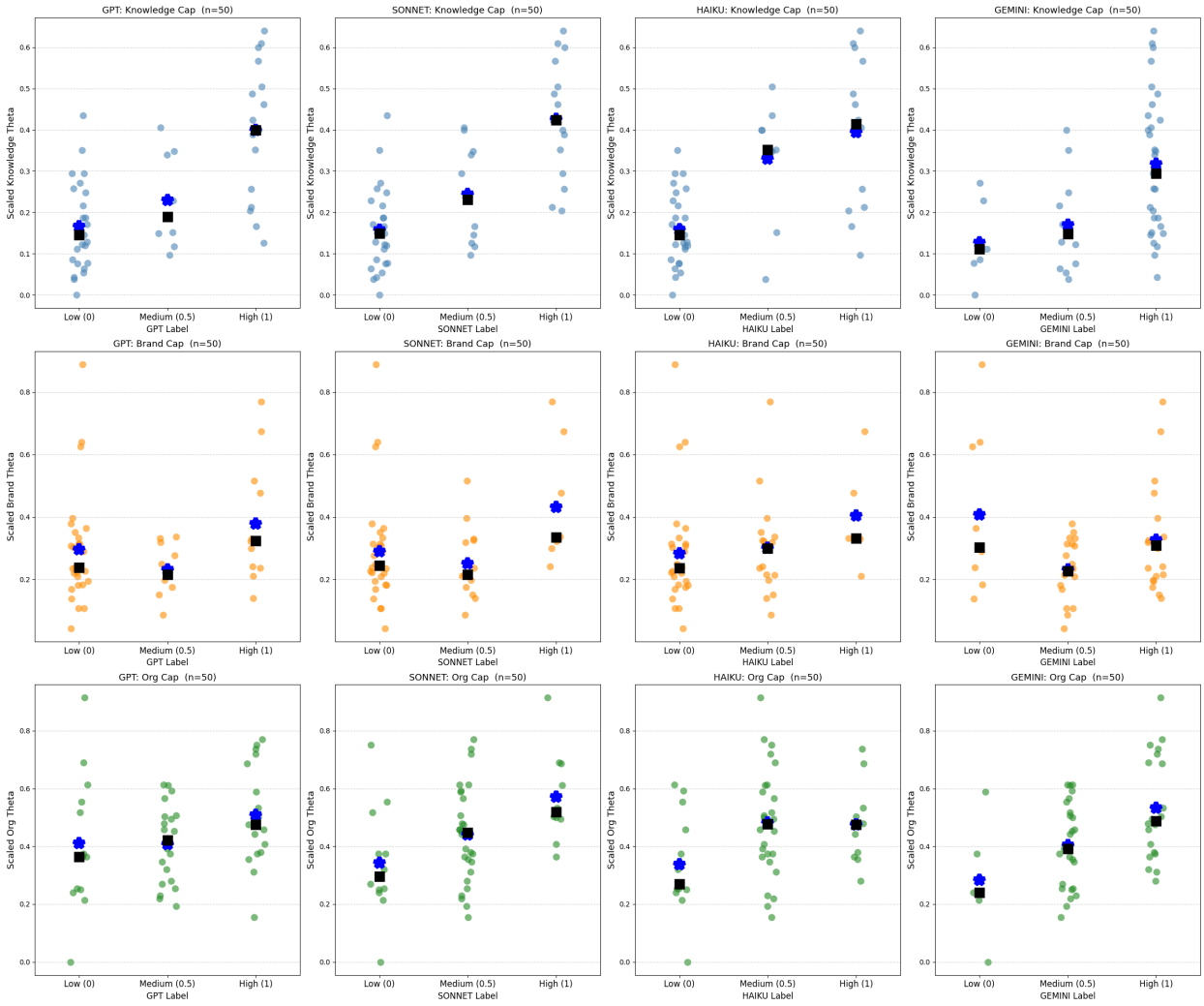


Figure A.1: STS Scores vs. Direct LLM Ratings

Notes: Each panel plots the STS θ score (y-axis) against the categorical rating assigned by a frontier LLM (x-axis: Low, Medium, High) for 50 10-K Item 7 documents from fiscal year 2023. Columns correspond to the three intangible capital components; rows correspond to GPT, Claude Sonnet, Claude Haiku and Gemini. Red and black markers indicate group means.

Across all nine panels, the relationship is positive: firms that receive higher LLM ratings tend to have higher θ scores, confirming that both approaches capture similar underlying signals about intangible investment intensity.

However, the direct LLM scoring approach exhibits substantial measurement noise. Table A.4 summarizes pairwise disagreement rates across 150 category-level ratings (50 documents $\times$ 3 components). Cross-model disagreement between Sonnet and Haiku ranges from 28.7% to 30.0%, but within-model disagreement, obtained by running Haiku twice on identical documents, is 26.0%. These rates are statistically indistinguishable, which implies that the majority of disagreement reflects noise intrinsic to the classification task and rubric ambiguity rather than meaningful differences across models. Organization capital is the noisiest component, with disagreement rates

Table A.4: LLM Scoring Disagreement Rates

Comparison	Overall	Knowledge	Customer	Organization
Sonnet vs. Haiku Run 1	28.7%	28.0%	18.0%	40.0%
Sonnet vs. Haiku Run 2	30.0%	30.0%	26.0%	34.0%
Haiku Run 1 vs. Run 2	26.0%	24.0%	16.0%	38.0%
Haiku Run 1 vs. Run 2 (temp = 0)	22.7%	20.0%	14.0%	34.0%

Notes: Each row reports the fraction of 150 category-level ratings (50 documents × 3 components) on which two scoring runs disagree. The first three rows use temperature = 1.0; the last row uses temperature = 0.0. Within-model disagreement (Haiku vs. Haiku) is statistically indistinguishable from cross-model disagreement (Sonnet vs. Haiku).

of 34–40% across all comparisons. Setting the sampling temperature to zero provides only modest improvement, reducing overall within-model disagreement from 26.0% to 22.7%.

The cost implications further favor our approach. Scoring 50 documents costs $5.84 with Sonnet and $1.55 with Haiku. Extrapolating to the full corpus of approximately 100,000 firm-year observations implies costs of roughly $11,700 and $3,100, respectively, per single run. Because the noise floor exceeds 20% even at temperature zero, reliable measurement would require multiple runs and an aggregation procedure, further multiplying costs. By contrast, STS scoring is deterministic: once the dictionary is constructed, each document receives the same score on every evaluation, at zero marginal cost.

These results highlight a fundamental trade-off. Direct LLM scoring applies richer contextual reasoning to each document but introduces stochastic noise that cannot be eliminated through model selection or temperature tuning. STS front-loads the semantic sophistication into dictionary construction and keeps the scoring step simple, fast, and fully reproducible. The positive correlation between the two approaches (Figure A.1) confirms that they capture similar economic signals, while the disagreement analysis demonstrates that STS offers superior reliability for large-scale measurement.

A.4 LLM Scoring Prompt for STS Validation

To benchmark the STS methodology against direct LLM scoring, we prompt frontier language models to classify the intangible investment intensity discussed in each firm's 10-K Item 7 text. The prompt instructs the model to identify all passages that discuss SG&A or operating expenses, categorize the discussed expenses into knowledge, customer, and organization capital, and assign a level of low, medium, or high for each component based on the depth and prominence of the discussion. Each model receives the identical prompt with the full Item 7 text appended. The prompt is reproduced below.

```
Below is the Item 7 (Management's Discussion and Analysis) section
from a company's 10-K filing. Your task is to assess the intensity
of intangible investment discussed in the company's SG&A and
operating expenses.

CATEGORY DEFINITIONS:

Intangible investments are expenses that build long-lived intangible
```

```
assets for the firm. They fall into three categories:

1. Knowledge capital: Spending that creates or advances proprietary
   knowledge, technology, or intellectual property. Examples include
   research and development expenses, R&D personnel costs, product
   candidate development, clinical trials, preclinical studies,
   laboratory supplies, equipment, CRO fees, technology development,
   engineering, process development, patent and IP-related activities,
   and scientific and technical personnel.

2. Brand or customer capital: Spending that builds brand recognition,
   customer relationships, or market position. Examples include
   advertising and promotion expenses, sales and marketing personnel
   and programs, customer acquisition costs, brand development, store
   brand marketing, commercialization activities, and distribution
   and sales channel development.

3. Organization capital: Spending that builds organizational
   capabilities, human capital, or operational processes. Examples
   include employee training and development programs, compensation
   and benefits designed to attract and retain talent (stock-based
   compensation, performance bonuses), payroll and employee benefit
   expenses tied to capability-building, management systems,
   organizational restructuring for efficiency, and IT systems and
   infrastructure for internal operations.

INSTRUCTIONS:

1. Read the entire document carefully.
2. Identify every passage that discusses, defines, or explains
   changes in SG&A expenses (or components thereof, such as
   "general and administrative expenses," "selling expenses,"
   "research and development expenses," "operating expenses").
3. For each expense-related passage, note which intangible
   investment category (if any) it relates to and how much detail
   is provided. Record this analysis in the "reasoning" field of
   your output BEFORE assigning levels.
4. Only after completing your passage-by-passage analysis, assign
   a level for each of the three intangible investment categories
   based on the evidence you gathered.

LEVEL DEFINITIONS:

- low: The category is absent, barely mentioned, or is a minor
  part of the discussion.
- medium: The category is discussed in moderate detail. It receives
  its own paragraph or several sentences explaining what it includes,
  why costs changed, or what activities it supports.
- high: The category is a major focus of the expense discussion.
  It is discussed at length, across multiple passages, with detail
  on specific programs, cost drivers, or strategic importance.

IMPORTANT NOTES:
- Focus only on passages that discuss specific expense categories.
  Ignore all other discussion in Item 7 (revenues, liquidity,
  capital resources, market conditions, forward-looking statements).
- Focus on the substance of what is described, not just the
  line-item labels.
- Assess each category independently.

OUTPUT FORMAT:

Return a single JSON object:

{
  "company_description": "<One-sentence description>",
  "sga_discussion_present": true,
  "reasoning": "<Summarize each expense-related passage you found,
   what category it maps to, and how much detail was provided.>",
  "levels": {
    "knowledge_capital": "<low|medium|high>",
    "brand_or_customer_capital": "<low|medium|high>",
    "organization_capital": "<low|medium|high>"
  }
}

DOCUMENT:

{document_text}
```

A.5 Calibration Using Tobin's q

An alternative calibration recovers the multipliers from valuations rather than from spending. We project Tobin's q, measured with physical capital in the denominator, on the three raw component stocks, with industry-group-by-year fixed effects and the multipliers constrained to be nonnegative. Under this normalization the scaled stocks enter the projection with a coefficient of one by construction, so the estimated slopes are directly comparable to the spending-share multipliers.

On the pooled sample the valuation-based multipliers are 2.75 for knowledge capital and 4.15 for customer capital, close to the spending-share estimates of 2.98 and 4.69, but 5.23 for organization capital, well above the spending-share estimate of 2.04. Estimated within individual industry groups, the valuation-based multipliers are unstable. The organization multiplier ranges from 3.4 to 9.7 across groups, and the knowledge multiplier is constrained at zero in one group, where the unconstrained estimate is negative.

We prefer the spending-share calibration for three reasons. It is based on observed expenditure rather than on market prices, which the valuation results in the paper treat as the outcome to be explained. It is far more stable across industry groups. And two of our three calibration approaches agree on an organization multiplier of approximately two. A formal treatment of the relationship between the two calibrations is in progress.

B Additional Figures and Tables

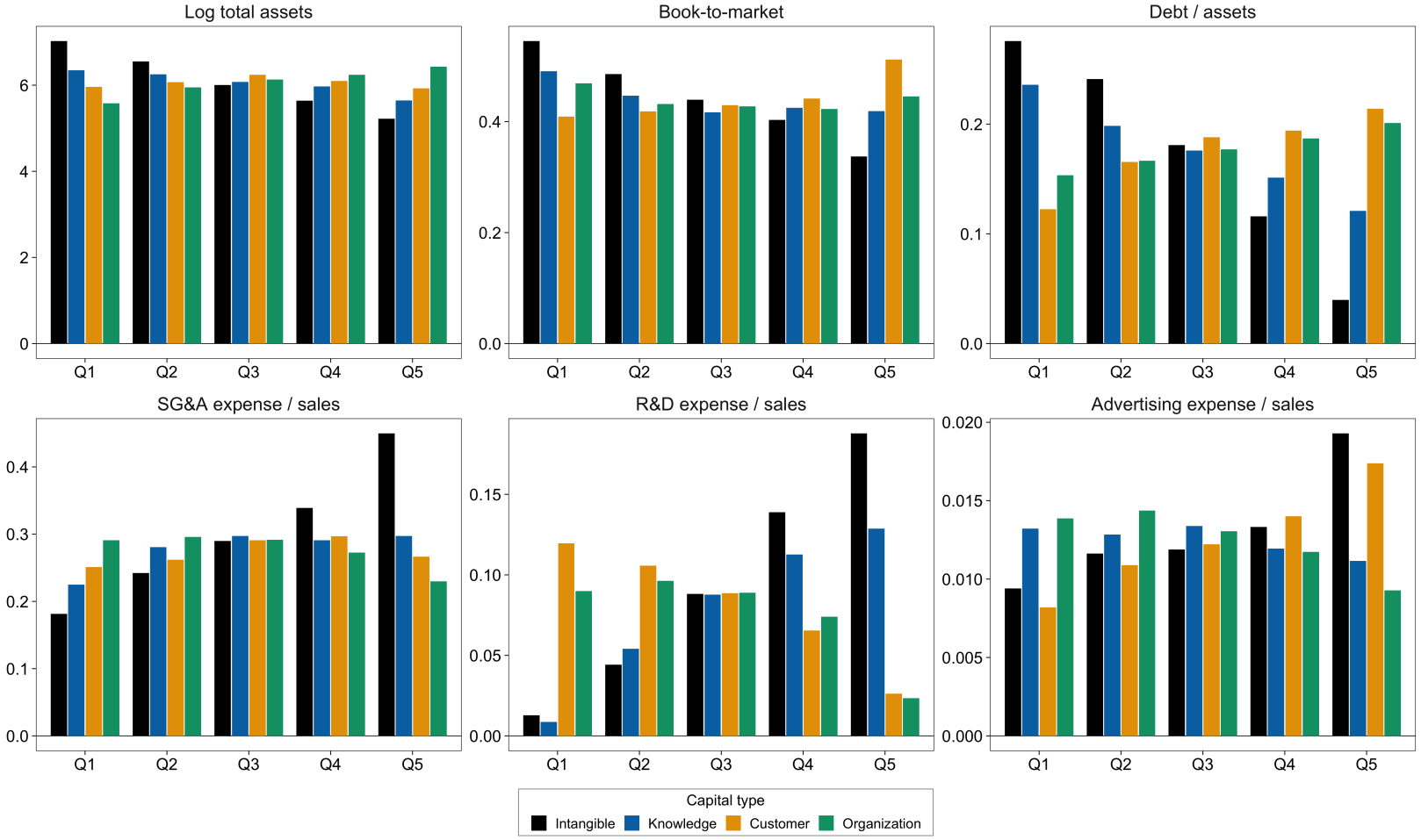


Figure B.1: Time-Averaged Values Across Quintiles

Notes: This figure plots time-averaged median firm characteristics across quintiles of firms sorted on θ^{int}, θ^{know}, θ^{cust}, and θ^{org}. Each panel reports one characteristic: the logarithm of total assets, the book-to-market ratio, debt to assets, and SG&A, R&D, and advertising expense over sales. Quintiles are assigned within industry-year cells as in Section 4.1. Characteristics are winsorized at the 1% level, medians are computed within year-quintile cells pooled across industries, and the yearly medians are averaged over 2002–2023.

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

Notes: This table reports five example communities from the dictionary-construction step of Section 2. Each community is a group of related n-grams identified by applying the Leiden community detection algorithm to the graph of K-Means micro-clusters. community_id is the community's identifier. size is the number of member n-grams. mean_similarity is the average pairwise cosine similarity among the embeddings of the community's member n-grams. representatives lists the community's highest-frequency member n-grams. category_name reports the community-level classification (intangible investment, not intangible investment, or unknown) assigned by an LLM and reviewed by the authors. subcategory_name assigns intangible-investment communities to knowledge, customer, or organization capital.

Table B.2: Industry-Level Summary Statistics of Intangible Intensity

Industry	Type	Mean	SD	P25	P50	P75
Food & Consumer Staples	θ^{int}	0.124	0.048	0.092	0.117	0.150
	θ^{cust}	0.420	0.160	0.303	0.412	0.525
	θ^{know}	0.162	0.118	0.082	0.138	0.214
	θ^{org}	0.412	0.164	0.293	0.407	0.528
Mining & Minerals	θ^{int}	0.093	0.047	0.060	0.081	0.112
	θ^{cust}	0.285	0.177	0.156	0.257	0.397
	θ^{know}	0.170	0.124	0.082	0.148	0.228
	θ^{org}	0.535	0.202	0.400	0.535	0.679
Oil & Petroleum Products	θ^{int}	0.078	0.036	0.056	0.071	0.092
	θ^{cust}	0.258	0.137	0.163	0.239	0.329
	θ^{know}	0.218	0.123	0.130	0.206	0.290
	θ^{org}	0.521	0.170	0.415	0.525	0.641
Textiles, Apparel & Footwear	θ^{int}	0.110	0.043	0.080	0.104	0.132
	θ^{cust}	0.426	0.146	0.325	0.417	0.521
	θ^{know}	0.166	0.103	0.092	0.147	0.219
	θ^{org}	0.408	0.157	0.303	0.405	0.506
Consumer Durables	θ^{int}	0.128	0.059	0.084	0.117	0.160
	θ^{cust}	0.345	0.152	0.238	0.327	0.445
	θ^{know}	0.239	0.128	0.149	0.228	0.326
	θ^{org}	0.408	0.162	0.297	0.400	0.519
Chemicals	θ^{int}	0.104	0.067	0.064	0.084	0.119
	θ^{cust}	0.283	0.142	0.185	0.262	0.368
	θ^{know}	0.232	0.140	0.127	0.217	0.315
	θ^{org}	0.469	0.168	0.361	0.468	0.587
Healthcare	θ^{int}	0.262	0.106	0.179	0.266	0.346
	θ^{cust}	0.182	0.134	0.074	0.150	0.263
	θ^{know}	0.457	0.206	0.305	0.481	0.627
	θ^{org}	0.359	0.164	0.246	0.329	0.435
Construction & Construction Materials	θ^{int}	0.099	0.043	0.072	0.092	0.121
	θ^{cust}	0.314	0.153	0.213	0.296	0.400
	θ^{know}	0.173	0.110	0.096	0.154	0.230
	θ^{org}	0.506	0.179	0.390	0.514	0.629
Steel Works Etc.	θ^{int}	0.076	0.035	0.053	0.071	0.099
	θ^{cust}	0.319	0.185	0.211	0.294	0.408
	θ^{know}	0.177	0.129	0.094	0.158	0.243
	θ^{org}	0.482	0.214	0.365	0.493	0.613
Fabricated Products & Machinery	θ^{int}	0.130	0.060	0.085	0.117	0.166
	θ^{cust}	0.283	0.124	0.196	0.268	0.350
	θ^{know}	0.300	0.144	0.194	0.297	0.390

Continued on next page

Industry	Type	Mean	SD	P25	P50	P75
Technology & Communications	θ^{org}	0.414	0.153	0.307	0.411	0.517
	θ^{int}	0.189	0.075	0.131	0.186	0.240
	θ^{cust}	0.300	0.129	0.209	0.286	0.375
	θ^{know}	0.308	0.145	0.205	0.301	0.403
	θ^{org}	0.391	0.145	0.294	0.381	0.477
Automobiles	θ^{int}	0.115	0.055	0.079	0.102	0.140
	θ^{cust}	0.307	0.135	0.208	0.299	0.391
	θ^{know}	0.258	0.143	0.150	0.254	0.354
	θ^{org}	0.435	0.150	0.339	0.429	0.529
Transportation	θ^{int}	0.100	0.047	0.069	0.091	0.120
	θ^{cust}	0.312	0.136	0.218	0.296	0.400
	θ^{know}	0.166	0.122	0.075	0.141	0.235
	θ^{org}	0.513	0.175	0.392	0.519	0.640
Retail Stores	θ^{int}	0.111	0.047	0.079	0.103	0.134
	θ^{cust}	0.385	0.158	0.270	0.376	0.488
	θ^{know}	0.125	0.098	0.056	0.105	0.167
	θ^{org}	0.483	0.172	0.368	0.483	0.602
Travel & Entertainment	θ^{int}	0.114	0.053	0.076	0.101	0.139
	θ^{cust}	0.351	0.142	0.246	0.333	0.433
	θ^{know}	0.173	0.110	0.091	0.162	0.229
	θ^{org}	0.476	0.149	0.377	0.476	0.580
Other	θ^{int}	0.131	0.069	0.082	0.115	0.167
	θ^{cust}	0.293	0.157	0.179	0.273	0.389
	θ^{know}	0.194	0.143	0.091	0.166	0.267
	θ^{org}	0.508	0.185	0.382	0.502	0.637

Notes: This table reports the mean, standard deviation, and quartiles of intangible intensity by industry, pooled across firm-year observations with non-missing interpolated scores, 2002–2023. For each industry, the θ^{int} row describes the level of the total score, while the θ^{cust}, θ^{know}, and θ^{org} rows describe each component's share of θ^{int}; the three shares sum to one for each firm-year. Utilities and Financials are absent from the estimation sample.

Table B.3: Intangible Intensity Rankings for Large Firms

Company	Ticker	SG&A/Sales	Med(SG&A)	R&D/Sales	Med(R&D)	Adv/Sales	Med(Adv)	θ_{int}	θ_{know}	θ_{brand}	θ_{org}
APPLE INC	AAPL	13.0%	48.6%	6.7%	16.6%		2.0%	1	5	1	2
MICROSOFT CORP	MSFT	26.3%	48.6%	12.4%	16.6%	0.8%	2.0%	2	5	2	2
ALPHABET INC	GOOG	29.0%	48.6%	14.0%	16.6%	3.3%	2.0%	4	3	3	3
AMAZON.COM INC	AMZN	41.2%	25.5%	14.2%	0.0%	4.0%	2.4%	4	5	4	1
NVIDIA CORP	NVDA	36.0%	48.6%	27.2%	16.6%		2.0%	2	5	1	1
JOHNSON & JOHNSON	JNJ	41.1%	84.6%	16.2%	69.8%	2.2%	2.2%	1	2	4	4
TESLA INC	TSLA	8.6%	18.4%	3.8%	4.2%		1.2%	4	4	2	3
WALMART INC	WMT	20.2%	25.5%	0.0%	0.0%	0.7%	2.4%	1	1	5	1
ELI LILLY & CO	LLY	47.8%	84.6%	28.4%	69.8%	3.5%	2.2%	1	4	1	3
PROCTER & GAMBLE CO	PG	25.1%	21.5%	2.5%	0.7%	9.9%	2.4%	2	3	4	3
HOME DEPOT INC	HD	16.7%	12.1%	0.0%	0.8%	0.7%	0.6%	1	1	1	5
PFIZER INC	PFE	24.5%	84.6%	12.5%	69.8%	2.8%	2.2%	2	3	4	3
ORACLE CORP	ORCL	38.5%	48.6%	17.3%	16.6%		2.0%	4	3	4	3
ABBVIE INC	ABBV	32.1%	84.6%	12.4%	69.8%	3.4%	2.2%	1	4	3	2
MERCK & CO INC	MRK	36.9%	84.6%	22.9%	69.8%	3.7%	2.2%	1	3	4	2
COCA-COLA CO	KO	30.0%	21.5%		0.7%	10.0%	2.4%	3	3	2	4
META PLATFORMS INC	META	50.9%	48.6%	30.3%	16.6%	2.3%	2.0%	3	3	2	4
PEPSICO INC	PEP	39.1%	21.5%	0.9%	0.7%	4.1%	2.4%	3	4	2	4
COSTCO WHOLESALE CORP	COST	8.7%	25.5%	0.0%	0.0%		2.4%	2	3	4	3

Notes: This table reports within-industry quintile ranks of θ^{int} and its three components for the largest firms by market capitalization in 2022, together with the underlying spending ratios. Ranks run from 1 (lowest) to 5 (highest) and are assigned within industry-year cells as in Section 4.1. SG&A/Sales, R&D/Sales, and Adv/Sales are the firm's expense-to-sales ratios; Med(·) columns report the corresponding industry-year medians. Blank cells denote missing Compustat expense items.

Table B.4: Spending-Share Calibration Multipliers by Industry Group

Industry group	m^{know}	N	m^{cust}	N	m^{org}	N
Consumer	3.00	992	4.97	802	2.19	4,514
Industrials	3.84	3,874	4.16	1,844	2.08	6,141
Healthcare	2.71	6,824	4.16	3,008	1.80	4,431
Technology	3.38	9,665	4.93	7,388	2.30	6,992
Transportation, Utilities & Other	3.07	1,704	4.48	978	1.75	4,713
Pooled	2.98	23,059	4.69	14,020	2.04	26,791

Notes: This table reports spending-share multipliers estimated separately within five broad industry groups, alongside the pooled estimates used in the paper. Each cell is a firm-year regression through the origin of an observed spending share on the corresponding component θ score within the group. Knowledge and customer spending shares are Capital IQ R&D and selling-and-marketing expenditure scaled by the sum of R&D and SG&A, with items reported as zero treated as missing, and the organization outcome is the equity-pay ownership share. All variables are winsorized at the 1st and 99th percentiles within fiscal year. The groups aggregate the paper's seventeen industries, and each group contains between 16% and 26% of firm-years. N is the number of firm-years in each regression. Standard errors are clustered by firm, and all estimates are significant at the 1% level.