Putting Marginal Back in Tobin's q*

Allen Hu† — *UBC*

Song Ma‡ — *Yale and NBER*

Preliminary & Incomplete — Please Do Not Circulate

Abstract

Marginal q—the shadow price of capital—is a sufficient statistic for investment, yet it is unobservable. For fifty years, empirical work has relied on the average Q, a weak predictor of investment. We take a bottom-up approach to seriously put "marginal" back to the q construction. For each firm-year, we use retrieval-augmented generation (RAG) to assemble relevant textual evidence from filings, patents, earnings calls, peer disclosures, and news, then prompt a large language model to identify plausible marginal projects and estimate their market values and costs. The resulting measure, q^{AI}, substantially outperforms average Q and intangible-adjusted total Q in predicting investment. Because we capture the marginal project itself, we show that the investment-q elasticity is larger for projects compatible with existing operations and smaller for acquisition-driven expansion. Decomposing the average Q-marginal q wedge, we find that capitalized rents, omitted intangibles, and financial frictions each contribute, with relative magnitudes varying across industries.

Keywords: Marginal q, Investment, Large Language Models, Retrieval-Augmented Generation, Intangible Capital

JEL Classification: G31, E22, C55

*First version: April 2026. This version: August 2026. We are grateful for the comments from Paul Goldsmith-Pinkham, Kyle Jensen, Kelly Shue, the workshop participants at CUHK Shenzhen, HKU, Michigan (Ross), UIUC (Gies). Eric Gao provided excellent research assistance. All remaining errors are ours.

†University of British Columbia, Sauder School of Business, email: allen.hu@sauder.ubc.ca.

‡Yale School of Management and NBER, email: song.ma@yale.edu.

1 Introduction

It is one of the most classic and central goals for economists to explain corporate investment. At the core of this pursuit is the q-theory of investment. Marginal q—the shadow value of one additional unit of capital—is the celebrated insight from investment theory dating back at least to Tobin (1969). When a firm faces a marginal investment opportunity, marginal q encodes all that matters for the decision: it is the sufficient statistic for explaining corporate investment (Hayashi, 1982; Abel and Blanchard, 1986; Caballero, 1999). A firm should invest whenever the market value of the incremental project exceeds its replacement cost—that is, whenever $q > 1$.

This theoretical elegance faces a fundamental empirical hurdle: we do not observe marginal investment opportunities, and thus we do not observe marginal q. For fifty years, empirical work has relied on a proxy: average Q, the ratio of a firm's market value to the replacement cost of its aggregate capital stock. The substitution of average for marginal is valid only under stringent assumptions—interior optimality, perfect competition, constant returns to scale in production, and adjustment costs that are homogeneous of degree one in capital (Hayashi, 1982)—and these conditions fail routinely. The empirical reality is unsatisfying: average Q is a weak predictor of investment, leaving most of the variation in investment rates unexplained.

Two main strategies have emerged to obtain a better-performing q measure. The first improves measurement within the average-Q framework: incorporating intangible assets into the capital stock (Peters and Taylor, 2017), using bond market prices to complement equity valuation (Philippon, 2009), or applying econometric corrections for classical noise (Erickson and Whited, 2000, 2012; Erickson, Jiang, and Whited, 2014). These approaches reduce error but accept the fundamental limitation that average Q is not marginal q. The second path estimates marginal q directly after assuming a static form of backward-looking production technologies. For example, Gala, Gomes, and Liu (2025) project manufacturing firms' market value onto measurable state variables—including the capital stock—then differentiate the fitted value function with respect to capital. The approach still relies entirely on the same Compustat balance-sheet variables that define average Q. The resulting marginal q is therefore a nonlinear transformation of accounting aggregates.

In this paper, we reverse the traditional approach. We ask: What if we could characterize the marginal projects, and build the q from there? The marginal investment project has never been

directly observed—it is a latent economic object. This is the fundamental reason the literature has been stuck with the average Q, sustained by a measurement paradigm that represents firms through accounting aggregates alone. We invert this logic. Instead of starting with total firm value and trying to infer the marginal from the average (top-down), we start by identifying a plausible set of investment opportunities a firm may face in a given year and construct q from there (bottom-up).

How can we systematically characterize a firm's plausible investment opportunities in each year? We rely on two things: vast quantities of text, and the modern generative AI method built on the idea of using existing information to predict the next thing. Text about firms are produced every day—strategic plans in 10-K filings, forward guidance in earnings calls, technical claims in patents, competitive assessments in peers' disclosures, macroeconomic context in news articles—and this text encodes granular and rich economic information that accounting numbers cannot. The text is so vast and heterogeneous that it raises the question of how to extract the right information from the right text at the right time for each firm. This is a retrieval problem, and retrieval-augmented generation (RAG) provides a principled solution: rather than assigning a fixed corpus to each observation, RAG searches across a heterogeneous universe of documents and assembles the most relevant pieces for the economic quantity of interest. We then use a large language model to propose plausible marginal projects and estimate their economic values, yielding a text-based measure of marginal q that we denote q^{AI}. The methodology proceeds in three steps.

- **Retrieval.** First, for each firm-year, we retrieve relevant passages from five text corpora—SEC 10-K filings, same-industry peer disclosures, patent filings, earnings conference call transcripts, and Wall Street Journal articles—subject to a best-effort no-look-ahead constraint that uses only information available prior to the focal year. Retrieval proceeds sequentially, with each stage conditioning on evidence already collected: the firm's own filings establish its strategic position, peer disclosures reveal competitive context, and patents, conference calls, and news fill in the innovation, operational, and macroeconomic dimensions.

- **Generation.** Second, we compress the retrieved evidence and prompt the LLM to propose three plausible next investment projects. The marginal project is not a scaled-down version of existing operations. It may involve entering a new market, acquiring a competitor, expanding physical capacity, or pursuing a new line of R&D—each with distinct value implications and

adjustment costs. Note that our goal is simply to characterize a firm's marginal investment opportunities; we do not necessarily need to fully predict the next project. All we require is that the characterization is *informative* about real investment opportunities.

- q **Construction.** Third, for each generated project, we produce an estimated market value and implementation cost. We compute project-level q as the ratio of market value to implementation cost and aggregate to a firm-year measure q^{AI}.

This method shows appealing properties when cross-checked against the data. We measure the similarity between the LLM-generated project descriptions and the firm's *future* 10-K filings and patent texts, and find that generated projects are significantly more similar to what the firm actually does in subsequent years than to random text. The similarity increases monotonically with q^{AI}: firms with higher AI-measured investment opportunities go on to describe and pursue projects more similar to those the system proposed. Decomposing the sources of variation in q^{AI} using a Shapley-Owen R^2 approach, we find that text carries more than 80 percent of q^{AI}'s variation; Compustat financials alone explain less than 18 percent. The composition varies meaningfully by industry—the firm's own 10-K dominates for manufacturing, patents matter most in high-tech, and conference calls contribute more in services—providing direct evidence that the measure draws on economically relevant textual content rather than simply repackaging accounting data.

A natural concern is look-ahead bias: the LLM may "know" what actually happened next, either through retrieved text or through its parametric memory. We address both channels. Text leakage is controlled by strict temporal filtering at the vector-database level—only documents dated before the focal year are eligible for retrieval. Parametric memory—the possibility that the LLM "remembers" realized outcomes from its training data—is harder to fully rule out. We mitigate it through evidence grounding via RAG and near-zero temperature decoding, which anchor the model's output in the retrieved evidence rather than background knowledge. We further test this concern by anonymizing firm identities in the evidence bundle, removing the cue that could trigger recall of specific corporate histories; the results are robust to this anonymization.

We use q^{AI} to revisit the core investment-q relation. In baseline panel regressions with firm and year fixed effects, q^{AI} substantially outperforms both standard average Q and intangible-adjusted total Q (Peters and Taylor, 2017) in predicting firm-level investment rates. This gain persists

in specifications controlling for cash flow, and the cash flow coefficient shrinks substantially—consistent with the prediction of Abel (2018) that better measurement of q drives out the spurious explanatory power of cash flow.

Applying the Erickson, Jiang, and Whited (2014) higher-order cumulant estimator, the measurement-error-corrected q^{AI} coefficients are substantially larger than OLS estimates, confirming that our measure retains less attenuation bias than conventional proxies. Extending the analysis to aggregate dynamics following Andrei, Mann, and Moyen (2019), we show that q^{AI} tracks post-2000 capital formation better than existing measures, capturing the decline in investment intensity without requiring ad-hoc adjustments. These results also bear on the longstanding financial constraints debate (Fazzari, Hubbard, and Petersen, 1988; Kaplan and Zingales, 1997)—whether cash flow effects in investment regressions reflect true financing frictions or measurement error in q—a question we revisit in our wedge decomposition in Section 5.

Our second set of results tests a refined prediction of investment theory: that the q-investment relationship should depend on the *nature* of the marginal project. Because we observe the marginal project—not just its value—we can test decompositions that aggregate measures cannot provide. Following Jovanovic and Rousseau (2014), we measure the extent to which each LLM-identified project is compatible with the firm's existing operations (intensive margin) versus requiring entry into new strategic territory (extensive margin). We find that the investment-q elasticity is substantially larger for intensive-margin projects than for extensive-margin ones, confirming that reallocation frictions compress the marginal response when firms must reorient their operations.

We further distinguish between "building" and "buying" on the margin—whether the marginal project involves organic capital expenditure or acquisition—following the q-theories of mergers in Jovanovic and Rousseau (2002) and Jovanovic and Rousseau (2003). The LLM's buy classification predicts actual M&A activity, and q^{AI} strongly predicts investment in both acquirer and non-acquirer subsamples, confirming that our measure captures genuine investment opportunities across implementation modes. This connects to the recent finding of Grullon, Ikenberry, and Verdi (2025) that excess capacity—anticipated underutilization of new capital—drives a wedge between average Q and the marginal return to investment. Our project-level decomposition complements their insight: we can directly observe what kind of project drives the marginal investment decision and how its characteristics shape the investment response.

Our final set of results conducts a residual analysis of the wedge between average Q and marginal q^{AI}, identifying several economically distinct sources: capitalized market-power rents, omitted intangible capital, scale effects, and financial frictions. The relative magnitudes vary systematically across industries and firm characteristics. In high-tech sectors, intangible omission dominates; in healthcare, both rents and intangibles contribute substantially; in manufacturing, the channels are more balanced.

This analysis extends the Q^+ framework of Crouzet and Eberly (2023), who show theoretically that the gap between valuations and investment reflects both rents from physical capital and the value of installed intangibles, and find that intangible-related terms contribute significantly, particularly in fast-growing sectors. Our contribution is to provide among the first firm-level empirical decompositions of this wedge, quantifying how much of the "investmentless growth" puzzle (Gutiérrez and Philippon, 2017) reflects market power versus omitted intangibles versus financial frictions. The analysis reveals that the two improvements pursued in prior work—econometric measurement correction and intangible capital adjustment—address only two of three problems and are, in combination, insufficient to recover marginal q reliably.

This paper's central contribution is to show that large language models, combined with structured retrieval over heterogeneous text corpora, can recover a bottom-up measure of marginal q that materially outperforms half a century of top-down accounting-based proxies. This opens a direct empirical window onto the marginal investment project—its nature, its adjustment costs, its realistic value—transforming q from an abstract shadow price into an observable, decomposable object. The implications extend across several literatures.

For the investment-q literature (Hayashi, 1982; Erickson and Whited, 2000; Peters and Taylor, 2017; Philippon, 2009; Gala, Gomes, and Liu, 2025; Andrei, Mann, and Moyen, 2019), we demonstrate that the longstanding weakness of average Q as a predictor is primarily a measurement problem, not a theoretical failure. For the measurement-of-q literature more specifically (Erickson, Jiang, and Whited, 2014; Gala, Gomes, and Liu, 2025; Grullon, Ikenberry, and Verdi, 2025; Belo et al., 2022; Ewens, Peters, and Wang, 2024), we move beyond accounting proxies entirely: our measure rests on textual evidence about the marginal project itself. For the financial constraints literature (Fazzari, Hubbard, and Petersen, 1988; Kaplan and Zingales, 1997; Hubbard, 1998; Abel, 2018), our project-level measure provides a sharper test of whether cash flow effects reflect

financing frictions or mismeasurement. For the rents-and-intangibles literature (Crouzet and Eberly, 2023; Gutiérrez and Philippon, 2017; De Loecker, Eeckhout, and Unger, 2020; De Ridder, 2024; Lindenberg and Ross, 1981), we provide among the first firm-level decompositions of the average-marginal wedge.

Our methodology also contributes to the growing literature that treats text as an input to economic measurement (Gentzkow, Kelly, and Taddy, 2019). Existing work uses text to *extract or classify* signals already encoded in documents—sentiment from corporate filings (Loughran and McDonald, 2011), product-market similarity from 10-K descriptions (Hoberg and Phillips, 2016), political risk from earnings calls (Hassan et al., 2019), earnings forecasts from financial statements (Kim, Muhn, and Nikolaev, 2024), and macroeconomic indicators from conference calls (Jha et al., 2024). Our paper takes a different step: we use text to *generate* a latent economic object—the marginal investment project—that no single document contains. The marginal project emerges only when evidence about the firm's capabilities, competitive environment, and macroeconomic context is assembled and synthesized. This is why the methodology pairs retrieval with generation: retrieval assembles relevant evidence from heterogeneous corpora, tailored to each firm-year, and the generative model synthesizes it into a project description, a market value, and an implementation cost. To our knowledge, this represents the first use of generative AI to construct, rather than extract or classify, a quantitative economic object.

The difficulty with measuring q is an instance of a broader challenge. Empirical economics has long relied on numerical information from balance sheets, income statements, and flow-of-funds tables as its primary measurement technology. This reliance is not specific to q: we measure profitability from earnings, risk from return volatility, intangible capital from capitalized R&D, and expected growth from analyst forecasts, each time accepting that structured numerical data are the best available summary of the underlying economic object. The retrieval-augmented generation pipeline we develop here—retrieve relevant text, compress, generate an economic quantity—is not specific to q either. Any latent economic object that leaves traces in firm text is a candidate: agency costs, financial constraints, risk premia, adjustment costs, expected growth, and managerial quality. Whether text-based measurement, constructed using the generative AI tools, can sharpen empirical tests across these domains is an open question that we hope this paper motivates.

The remainder of the paper is organized as follows. Section 2 describes the data sources

and sample construction. Section 3 details the LLM-RAG methodology—retrieval, compression, generation, and q construction—and validates the approach through forward-similarity tests and a Shapley-Owen decomposition of information sources. Section 4 presents the main investment-q results, including baseline regressions, cash flow controls, measurement error correction, and aggregate time-series evidence. Section 5 exploits the project-level content of q^{AI} to test heterogeneity by project type, compatibility with existing operations, and build-versus-buy implementation mode, and then decomposes the wedge between average Q and marginal q^{AI}. Section 6 concludes.

2 Data and Sample Construction

Measuring marginal q from the bottom up requires two categories of data: standard financial variables needed to measure investment and capital, and text corpora that encode forward-looking information about investment opportunities absent from accounting aggregates. This section describes both and details how we process unstructured text for economic analysis and construct our estimation sample.

2.1 Financial Data and Investment Variables

Our financial data come from standard sources. We obtain annual accounting variables from Compustat and stock market data from CRSP, merging the two via the WRDS CRSP-Compustat link table. Each observation is a firm-year, identified by Compustat's permanent firm identifier (`gvkey`) and fiscal year.

We construct the investment rate as the ratio of capital expenditures to the lagged capital stock, $I_{it}/K_{i,t-1}$, following the standard specification in the investment-q literature. The capital stock is measured as net property, plant, and equipment. We compute *average* Tobin's q as the market-to-book ratio of assets: the market value of the firm (equity market capitalization plus book value of liabilities) divided by book value of total assets. As a benchmark, we also construct the intangible-adjusted *total* q of Peters and Taylor (2017), which augments both numerator and denominator with estimates of knowledge capital (accumulated R&D) and organization capital (accumulated SG&A), accounting for the growing share of intangible investment omitted from traditional balance sheets.

We apply standard sample filters. We exclude financial firms (SIC codes 6000–6999) and

utilities (SIC codes 4900–4999), whose regulated capital structures make standard investment-q regressions difficult to interpret. We require positive book assets and non-missing data on capital expenditures, depreciation, and market capitalization. We limit the sample to firm-years with shares listed on the NYSE, NASDAQ, or AMEX (CRSP share codes 10 and 11) to ensure reliable market valuations.

2.2 Text Corpora

The key insight motivating our data strategy is that identifying the marginal investment project requires information about the firm's strategic direction, competitive position, innovation pipeline, and macroeconomic context—information that is encoded in corporate text and largely invisible in accounting aggregates. We assemble five text corpora that, taken together, reconstruct the informational landscape available to a firm's decision-makers at each point in time. Table 1 summarizes the corpora, their primary identifiers, and the descriptive fields we retain. Each corpus captures a distinct dimension of the investment opportunity set; we describe them in turn.

2.2.1 Corporate Annual Reports (SEC 10-K Filings)

Our primary text source is the universe of annual 10-K filings available through the SEC's EDGAR system. We extract two sections from each filing. *Item 1* ("Business") provides a comprehensive description of the firm's products, services, competitive position, and industry environment—a structured narrative of what the firm *is* at a given point in time. *Item 7* ("Management's Discussion and Analysis of Financial Condition and Results of Operations") is forward-looking by regulatory design: SEC rules require management to discuss known trends, events, and uncertainties reasonably expected to affect future financial performance.[1] Item 7 therefore encodes precisely the kind of forward-looking managerial assessment—planned expansions, anticipated competitive pressures, expected capital requirements—that is relevant for inferring marginal investment opportunities.

In the language of structural investment models, Item 1 identifies the firm's existing capital structure and productivity state—the "state" from which the marginal project departs—while Item 7

[1]The growing use of 10-K text in financial economics includes sentiment analysis (see, e.g., Loughran and McDonald, 2011), text-based industry classification (Hoberg and Phillips, 2016), and firm-level embedding representations (Sarkar, 2025).

reveals management's assessment of how that state is likely to evolve. Together, they provide the textual analog of what a structural model would need to distinguish marginal from average q: the current productive capacity and the expected direction of change.

2.2.2 Peer Firm Disclosures

The marginal investment project depends not only on internal capabilities but on competitive context. A competitor's capacity expansion alters the focal firm's residual demand curve and thereby the value of its marginal investment; a peer's entry into a new product market shifts the competitive landscape. We therefore retrieve 10-K filings (Items 1 and 7) from peer firms in the same industry, defined using the Standard Industrial Classification (SIC) system. We restrict retrieval to filings from prior years. Including peer disclosures allows the language model to condition its project proposals on the competitive environment the firm actually faces—a dimension of the investment opportunity set that the firm's own disclosures may understate or omit entirely.

2.2.3 Earnings Conference Call Transcripts

Annual 10-K filings are static snapshots; managerial expectations and strategic priorities evolve throughout the year. We therefore supplement annual disclosures with transcripts of quarterly earnings conference calls. These calls feature prepared remarks by senior management—typically the CEO and CFO—followed by a question-and-answer session with sell-side analysts. The Q&A format is particularly informative because it allows analysts to probe management on specific investment plans and strategic priorities, often eliciting more detailed responses than formal SEC filings.[2]

Conference calls are valuable for our purposes along two dimensions. First, their quarterly frequency captures shifts in managerial expectations between annual filings—for instance, a mid-year pivot toward acquisition-led growth or a newly announced capacity expansion. Second, the interactive format surfaces information that firms might not volunteer in regulatory filings, including candid assessments of competitive threats and forward-looking project-level detail. We retain transcripts occurring prior to the focal fiscal year to maintain temporal integrity.

[2]Conference call transcripts have become a workhorse data source in the text-as-data literature in finance. See, e.g., Bushee, Matsumoto, and Miller (2003), Matsumoto, Pronk, and Roelofsen (2011), and Hassan et al. (2019), who use conference call text to construct firm-level measures of political risk.

2.2.4 Patent Filings

To capture the innovation dimension of the investment opportunity set, we incorporate patent texts from the U.S. Patent and Trademark Office (USPTO). Each patent document contains a detailed technical description of the claimed invention, providing a window into the firm's R&D pipeline and technological direction. Patent filings reveal not just *that* a firm innovates, but *what* it innovates on—the specific technological domains and engineering challenges that define its innovation frontier. This is especially important because the marginal project in many technology-intensive firms involves intangible capital investment that balance sheet accounting systematically misses but patent filings directly capture.

We link patents to Compustat firms at the patent-year level, matching patent assignees to firm identifiers using established name-matching procedures and supplementary crosswalks. This matching must account for the fact that corporate assignee names on patents do not always correspond exactly to legal entity names in financial databases. We retain patent-level information including filing year and title. The economic content of patent data is well established: Kogan et al. (2017) estimate the private value of patents using stock market responses to patent grants, and Woeppel (2022) demonstrates that patent capital is a powerful predictor in the investment-q framework, directly motivating our inclusion of patent text.

2.2.5 Financial News: Wall Street Journal

The final corpus provides macroeconomic context. We collect front-page articles from the *Wall Street Journal* to capture aggregate economic, financial, and policy developments that affect the investment opportunity set. WSJ coverage has been used extensively in recent work on narrative economics: Bybee, Kelly, and Su (2023) demonstrate that text-derived factors from news articles carry meaningful information about systematic risk premia, and a large related literature uses newspaper text to construct measures of economic policy uncertainty.[3]

For our purposes, WSJ articles provide the macroeconomic backdrop within which firm-level investment decisions are made. Aggregate demand conditions and interest rate expectations determine the discount rates applied to marginal projects; trade policy developments and sector-wide

[3]See, e.g., Baker, Bloom, and Davis (2016), who construct a monthly policy uncertainty index from major newspapers including the WSJ.

technology shifts directly affect expected project cash flows. By including news articles in the retrieval system, we allow the language model to condition its project proposals on the same information environment that firm managers observed when making actual investment decisions. We index articles by year and restrict retrieval to articles published before the focal fiscal year.

2.3 From Text to Numerical Representations

For each firm-year, we face a retrieval problem: identifying, from millions of documents across five corpora, the specific textual evidence most relevant to the firm's investment opportunity set. The relevant evidence is scattered across the firm's own disclosures, its competitors' filings, its patent portfolio, management's oral commentary, and contemporaneous news—and the composition of this relevant set varies across firms and over time. This subsection describes the data infrastructure that makes such tailored retrieval feasible; Section 3 details the full retrieval and generation algorithms, and Appendix D provides the complete technical specification of the text processing pipeline.

Text embeddings. We convert each text passage into a fixed-length numerical vector—an "embedding"—such that passages with similar meaning are mapped to nearby points in vector space. In economic terms, this assigns each passage a numerical signature that captures its content: passages discussing related topics (e.g., "capacity expansion" and "new production facility") receive similar signatures, enabling the system to identify relevant evidence automatically. Formally, an embedding model maps each text passage C_k to a vector $\mathbf{v}_k \in \mathbb{R}^{1024}$, where cosine similarity approximates semantic similarity. This mirrors recent work on text-based economic representations (Bybee, Kelly, and Su, 2023; Gabaix et al., 2024; Sarkar, 2025): economic objects are positioned in a learned vector space where proximity encodes relatedness.

Document segmentation and storage. Because individual documents—a full 10-K Item 1 section, a complete conference call transcript—are too long to embed as single units, we split each document into overlapping text segments of approximately 512 characters. The overlap ensures that semantically coherent sentences spanning a segment boundary are captured. Each segment retains its source document's identifying information—firm identifier, year, source type—enabling

retrieval to be conditioned on economic attributes such as industry, firm identity, and time. Segments from all five corpora are stored in a vector index organized by source type, allowing retrieval to be restricted by corpus as needed.

Retrieval. For a given firm-year, the retrieval system identifies the most relevant text segments from each corpus by searching over the embedding space, subject to filters that enforce firm-level and temporal constraints. This design allows the system to assemble a bespoke evidence bundle for each firm-year—drawing from the firm's own history, its peers, its patents, its management's recent commentary, and the contemporaneous news environment—rather than applying a fixed, researcher-defined text assignment. The retrieval architecture supports multi-stage refinement: evidence from one corpus (e.g., the firm's 10-K) refines queries applied to others (e.g., peers and patents). Section 3 describes this sequential design in detail.

No-look-ahead constraint. A critical design choice is that all retrieved text must pre-date the focal fiscal year. For each firm-year (`gvkey`, `fyear`), we apply strict temporal filters that exclude any document dated in or after the focal year. This constraint is essential for valid empirical tests of the investment-q relation, because it ensures that our text-based q measure is constructed using only information available to the firm and its investors at the time of the investment decision. Using future earnings announcements or post-period filings to infer current investment opportunities would introduce look-ahead bias and overstate the explanatory power of q^{AI}.

2.4 Sample Construction

Our estimation sample is constructed by merging the financial panel from Compustat-CRSP with the text retrieval system. A firm-year enters the sample if it satisfies the financial data filters described in Section 2.1 and the retrieval system produces a non-empty evidence bundle—that is, the firm has sufficient prior text data from at least one corpus to support meaningful retrieval.

In practice, the binding constraint on sample coverage is the availability of text in early years: EDGAR coverage of 10-K filings becomes comprehensive in the late 1990s, and conference call transcripts are available primarily from the mid-2000s onward. We restrict the sample to firms with at least 25 years of observations in our panel, ensuring sufficient prior disclosures to support

meaningful retrieval.

This sample restriction introduces a selection effect that is worth discussing. The resulting sample is weighted toward mature, established firms with long histories of SEC filings and analyst coverage. This has both advantages and limitations. On the positive side, mature firms account for a disproportionate share of aggregate capital formation and have been the central focus of the investment-q literature since Hayashi (1982). Their investment behavior is therefore the natural benchmark for evaluating a new q measure. On the other hand, we likely miss young and rapidly growing firms whose marginal projects are arguably the most interesting from the perspective of capital allocation. We revisit the implications of this selection in the empirical analysis.

Table 1. Text corpora, primary identifiers, and stored fields.

Notes. This table summarizes the five text corpora used for retrieval-augmented generation. "Primary key" indicates the firm-time identifiers used to index and filter documents. All corpora are subject to the no-look-ahead constraint: only documents dated strictly before the focal fiscal year are retrieved. See Section 2.3 for details on the embedding and retrieval infrastructure.

Source	Primary Key	Stored Fields
10-K Item 1	(gvkey, fyear)	Company name; SIC industry code; CIK; filing date; form type; accession number; section title
10-K Item 7	(gvkey, fyear)	Same as Item 1; distinguished by section title
Patents	(gvkey, filing year)	Patent title; patent identifier
Conference calls	(CUSIP, fiscal year)	Latest call date
WSJ front pages	(year)	Article title

[Insert Table 2 Here.]

3 Methodology: From Text to Marginal q

Measuring marginal q from the bottom up requires a system that can, for each firm-year, identify the most likely next investment project and assign it an economic value. No single text source contains this information in isolation. A firm's strategic direction is encoded in its annual filing; its competitive context emerges from peer disclosures; the specifics of its innovation pipeline appear in patent texts; its near-term investment priorities surface in earnings calls; and the macroeconomic backdrop—interest rate expectations, trade policy, sector-wide demand conditions—shapes the

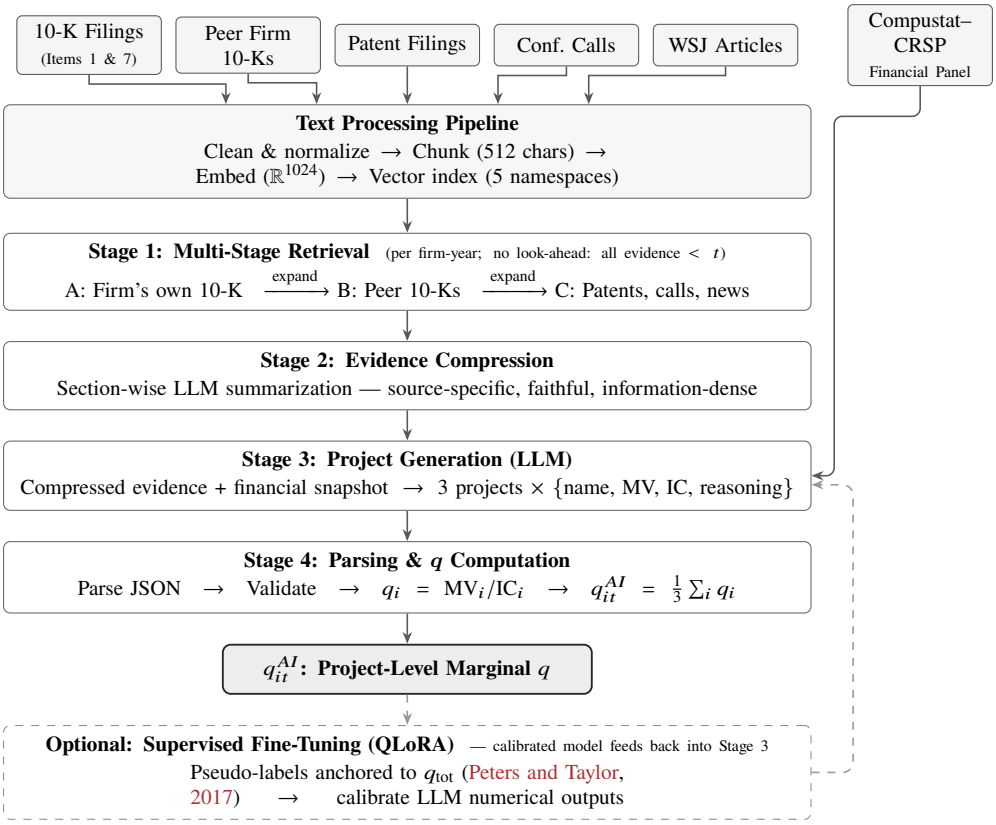


Figure 1. Pipeline for constructing project-level marginal q. Raw text from five corpora is cleaned, chunked, and embedded into a vector index (Section 2.3). For each firm-year, the pipeline retrieves relevant evidence in three sequential stages with query expansion (Section 3.1), compresses it into faithful source-specific summaries (Section 3.2), and generates three project proposals grounded in both textual evidence and financial data (Section 3.3). Project-level q is computed as the ratio of market value to implementation cost (Section 3.4). An optional supervised fine-tuning stage calibrates numerical outputs to an external benchmark (Section 3.5). All retrieved evidence predates the focal year.

discount rates and cash flow forecasts relevant to any marginal project. The methodological challenge is to assemble, from these heterogeneous sources, a tailored evidence bundle for each firm-year, and then to convert that evidence into a structured economic object: a proposed project with an estimated market value and implementation cost.

We address this challenge using retrieval-augmented generation (RAG), a framework in which a large language model (LLM) generates output conditioned not only on its parametric knowledge but also on evidence retrieved at inference time from an external corpus. The methodological advantage of RAG is that it isolates two tasks that are conceptually distinct: identifying which information is relevant to a particular firm's investment decision (retrieval), and reasoning over that information to propose and value a marginal project (generation). This separation mirrors how an analyst would approach the problem—first gathering the relevant filings, transcripts, and news, then synthesizing that material into an investment thesis.

The pipeline proceeds in five stages. First, we *retrieve* relevant text passages from five corpora, subject to a strict no-look-ahead constraint, assembling a multi-source evidence bundle for each firm-year (Section 3.1). Second, we *compress* this evidence into a concise, faithful summary that preserves provenance across sources (Section 3.2). Third, we *generate* project proposals by presenting the compressed evidence alongside a structured financial snapshot to the LLM and asking it to propose exactly three plausible next projects with estimated market values and implementation costs (Section 3.3). Fourth, we *parse* the structured output and compute project-level q as the ratio of market value to implementation cost (Section 3.4). Fifth, we optionally *calibrate* the model through supervised fine-tuning on pseudo-labels anchored to an external benchmark, improving the numerical discipline of the generated valuations (Section 3.5). Appendix E provides the complete technical specification of each stage, including algorithm pseudocode, hyperparameter configurations, and implementation details.

3.1 Retrieval: Assembling the Evidence Bundle

The retrieval stage maps each firm-year into a structured, multi-source evidence bundle. The economic motivation is straightforward: the marginal investment project depends simultaneously on the firm's internal capabilities, its competitive environment, its innovation frontier, and the

macroeconomic moment. No single text corpus captures all of these dimensions. Our retrieval system searches across the five corpora described in Section 2.2—the firm's own 10-K filings, same-industry peer disclosures, patent filings, earnings conference call transcripts, and Wall Street Journal articles—and selects the passages most semantically relevant to the firm's investment opportunity set.

Temporal integrity. All retrieval is subject to a strict historical constraint: for a given firm-year (i, t), only documents dated before year t are eligible. This no-look-ahead restriction is essential for valid empirical tests of the investment-q relation. Using future disclosures or post-period news to infer current investment opportunities would introduce look-ahead bias and overstate the explanatory power of q^{AI}. The constraint is implemented through metadata filters applied at the vector database level, ensuring that temporal integrity is enforced mechanically rather than relying on researcher judgment.

Sequential retrieval with query expansion. A key design choice is that retrieval proceeds sequentially across sources, with each stage conditioning on evidence already collected. We first retrieve the focal firm's own historical 10-K disclosures (Item 1, then Item 7), which establish the firm's current strategic position and management's forward-looking assessment. The retrieval query for Item 7 is expanded to include the already-retrieved Item 1 content, so that the system selects managerial commentary that complements—rather than duplicates—the business description. Next, we retrieve peer firm disclosures from the same industry, using a query that incorporates the focal firm's own retrieved filings. This sequential design ensures that peer evidence is selected for its relevance to the focal firm's specific situation, not merely for its semantic proximity to a generic industry query.

In the final stage, we retrieve from the remaining corpora—patents, conference calls, and WSJ articles—using a query that incorporates both the firm's own disclosures and its peers' filings. This progressive query expansion implements a form of pseudo-relevance feedback: the evidence retrieved at each stage refines the search at subsequent stages, producing a coherent evidence bundle in which the sources are complementary rather than redundant.

Semantic search with metadata filtering. Within each corpus, retrieval is performed by embedding the query into the same vector space as the stored text segments (Section 2.3) and identifying the nearest neighbors by cosine similarity, subject to metadata filters that enforce firm-level and temporal constraints. For the firm's own filings, we filter on the firm identifier; for peers, we filter on industry code while excluding the focal firm; for patents, on the firm identifier and filing year; for conference calls, on the CUSIP identifier; and for WSJ, on publication year. This combination of semantic similarity and metadata filtering allows the system to find evidence that is both topically relevant and temporally valid. Appendix E.1 documents the retriever implementation, including the deduplication heuristics, reranking procedure, and per-corpus retrieval budgets.

3.2 Evidence Compression

The raw evidence bundle from retrieval is typically redundant and exceeds the model's effective context window. Text segments retrieved from overlapping chunks of the same filing may repeat identical sentences; boilerplate language appears across multiple years of the same firm's disclosures; and the combined volume of evidence from all five corpora is too large for stable downstream generation.

We therefore introduce a post-retrieval compression layer that rewrites each source-specific block of evidence into a dense, faithful summary. The compression operates independently on each block—the firm's own filings, peer disclosures, patents, conference calls, and news articles—preserving the provenance boundary between sources. This source-specific design prevents facts from one corpus (e.g., peer disclosures) from contaminating summaries of another (e.g., the firm's own filings).

The summarization objective imposes two hard constraints. First, *faithfulness*: the summary must draw only on facts present in the original evidence block, with no invention or speculation. This constraint is critical because any hallucinated content injected at the compression stage would propagate into the downstream project proposals and valuations. Second, *information density*: the summary must preserve concrete, economically relevant details—dates, quantities, segment names, competitor mentions, product lines, guidance language—that would be useful for inferring marginal investment projects.

The summaries follow a standardized schema aligned with the analytical goals of the subsequent generation stage: core business and strategy, performance signals, current projects and investment focus, competitive landscape, and risks and outlook. This structured format ensures that the compressed evidence is directly usable for project inference without further reorganization. Appendix E.4 provides the full specification of the compression procedure, including token budgeting, overflow handling, and the section-marker scaffold.

3.3 Project Generation

The generation stage converts the compressed evidence bundle into a set of proposed marginal investment projects. This is the step at which the system moves from information retrieval to economic inference—from "what does the evidence say about this firm?" to "what is the firm most likely to invest in next, and what is that investment worth?"

Prompt construction. The generation prompt merges two types of information. The first is the compressed multi-source evidence from the preceding stage, which encodes the firm's strategic position, competitive environment, innovation pipeline, and macroeconomic context. The second is a structured financial snapshot for the focal fiscal year, drawn from the Compustat-CRSP panel, which provides the quantitative anchors—total assets, capital expenditures, R&D spending, market capitalization, revenue, and profitability—needed to ground the model's project valuations in the firm's actual scale.

The prompt instructs the model to propose exactly three plausible next investment projects for the firm. For each project, the model must provide: a project name, a detailed description, an estimated market value (in USD millions), an estimated implementation cost (in USD millions), a reasoning narrative explaining the economic logic, a confidence score, and three comparable public firms. The requirement of three projects—rather than a single best guess—captures the inherent uncertainty in identifying the marginal project and provides a basis for examining cross-project variation in q.

Structured output and large-scale inference. The model is constrained to return its response as a strict JSON array, enabling automated extraction of the numerical quantities needed to compute

project-level q. We generate outputs using conservative decoding (near-zero temperature) to maximize reproducibility. Generating proposals for the full firm-year panel requires processing thousands of long prompts; we implement this using a high-throughput serving framework with built-in fault tolerance and incremental checkpointing (Appendix E.5).

3.4 Computing Project-Level q

Having generated project proposals, we extract the key economic quantity of interest: project-level marginal q. For each proposed project $i \in \{1, 2, 3\}$, we compute the implied marginal q as:

$$q_i = \frac{\texttt{MARKET VALUE}_i}{\texttt{IMPLEMENTATION COST}_i}, \tag{1}$$

where both quantities are denominated in USD millions and are extracted from the model's structured JSON output. When the denominator is missing, zero, or non-numeric, we set q_i to missing.

This ratio has a direct economic interpretation. If the model identifies a project whose expected market value exceeds its implementation cost ($q_i > 1$), the project creates value and should, under standard investment theory, be undertaken. The magnitude of q_i reflects the intensity of the investment opportunity: a project with $q_i = 2$ is twice as valuable per dollar of implementation cost as one with $q_i = 1$. Because we observe three project proposals per firm-year, we can examine both the level and dispersion of q across a firm's marginal investment opportunities.

From projects to firm-year q^{AI}. Our baseline firm-year measure, q^{AI}, is the simple average of the three project-level ratios:

$$q_{it}^{AI} = \frac{1}{3} \sum_{i=1}^{3} q_i. \tag{2}$$

This averaging treats each proposed project as an equally likely candidate for the marginal investment. In the empirical analysis, we also examine results using individual project-level q values and alternative aggregation schemes to assess robustness.

Parsing and validation. We implement a deterministic parsing layer that recovers valid JSON arrays from outputs with minor formatting artifacts, and we validate that each output contains exactly three project objects with numerically valid market values and strictly positive implementation costs.

Only valid observations contribute to downstream analysis; we report the valid-output rate in the empirical results. The parsing procedure is documented in Appendix E.6.

[Insert Table 2 Here.]

3.5 Supervised Fine-Tuning with Pseudo-Labels

The direct RAG pipeline produces project proposals that are qualitatively rich—grounded in multi-source evidence, specific to the firm's situation, and responsive to the competitive and macroeconomic environment. However, the numerical outputs (market values and implementation costs) may be noisy, particularly for firms with unusual scale or in industries where the model has limited parametric knowledge. To improve the numerical discipline of the generated valuations while preserving the qualitative grounding provided by RAG, we introduce an optional supervised fine-tuning (SFT) stage.

3.5.1 Pseudo-Label Construction

The fundamental challenge in fine-tuning a model to produce project-level valuations is the absence of ground-truth labels: we do not observe the true market value or implementation cost of the marginal project. We address this by constructing pseudo-labels that anchor the model's numerical outputs to an observable firm-year benchmark while preserving the qualitative project structure generated by the direct RAG pipeline.

Specifically, we use the intangible-adjusted total q of Peters and Taylor (2017), denoted q_{tot}, as an external calibration target. For each firm-year, we take the model's original three-project output—with its project names, descriptions, reasoning, and comparable firms intact—and recalibrate only the numerical fields. Each project's implementation cost is set equal to total investment (capital expenditures plus R&D and SG&A investment) parsed from the financial snapshot, and the corresponding market value is then chosen so that:

$$q_i^{\text{label}} = \frac{\texttt{MARKET VALUE}_i^{\text{label}}}{\texttt{IMPLEMENTATION COST}_i^{\text{label}}} \rightarrow q_i^{\text{tot}}. \tag{3}$$

Each of the three projects receives an identical q label equal to q_{tot}, differing only in project names,

descriptions, and reasoning. This symmetric design reflects the fact that we do not observe project-level valuations directly; we use q_{tot} as a firm-year scalar anchor rather than a project-specific outcome. The resulting pseudo-label is not human-annotated ground truth; it is a programmatic supervision target that combines the model's own qualitative project structure with an externally observed valuation benchmark. Although the pseudo-labels assign the same q to all three projects, the model is trained on the full prompt context—compressed evidence plus financial snapshot—and must generate distinct project proposals with differentiated names, descriptions, and reasoning, preserving project-level heterogeneity in the outputs. The choice of q_{tot} as the anchor is motivated by its role as the best available accounting-based proxy for the overall firm-year valuation ratio (Peters and Taylor, 2017), providing a principled starting point for calibration.

3.5.2 Fine-Tuning Objective and Procedure

Each training example consists of the full RAG-grounded generation prompt (compressed evidence plus financial snapshot plus output instructions) paired with the pseudo-label completion (the recalibrated three-project JSON). We train the model using a completion-only objective:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{n=1}^{N} \sum_{t \in \mathcal{T}(y_n)} \log p_\theta(y_{n,t} \mid x_n, y_{n,<t}), \tag{4}$$

where x_n is the prompt, y_n is the pseudo-label completion, and $\mathcal{T}(y_n)$ indexes only the completion tokens. Tokens belonging to the prompt receive a loss mask, so the model is trained to generate the calibrated JSON output given the RAG-grounded input, rather than being penalized for reproducing the prompt itself.

We implement SFT using QLoRA, a parameter-efficient adaptation method that attaches low-rank adapters to the base model's attention and feedforward layers while keeping the base parameters frozen in 4-bit quantized form. This design allows fine-tuning on a single GPU with modest memory requirements, making the calibration step practical even for large instruction-tuned models. The default configuration uses LoRA rank $r = 16$ and scaling parameter $\alpha = 32$.

Why fine-tuning helps. The direct RAG pipeline excels at retrieving evidence and reasoning qualitatively about investment opportunities, but translating qualitative assessments into dollar-

denominated market values requires anchoring to observable economic magnitudes. Under zero-shot inference, the model's parametric priors about firm scale and discount rates may introduce systematic bias—for instance, overvaluing novel projects or underestimating implementation costs in capital-intensive industries. SFT adapts the model's numerical outputs to the domain of firm-year valuations, improving internal consistency with observed q_{tot} while preserving the qualitative project grounding provided by RAG.

3.5.3 Train-Test Design and Evaluation

We partition the pseudo-labeled sample using grouped random splits at the firm level to prevent information leakage across firm-years of the same company. The default partition is 80/10/10 (train/validation/test), ensuring that evaluation is conducted on held-out firms rather than held-out years of the same firm.

On the test set, we evaluate the adapted model by generating completions deterministically and comparing the implied project-level q values against the benchmark q_{tot}. We report four error metrics—mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and log-scale absolute error ($\text{MAE}_{\log}$)—along with the valid-JSON rate, defined as the fraction of test prompts for which the model returns a parseable three-project output.

This evaluation is a calibration test: it assesses whether the fine-tuned model produces numerically coherent outputs anchored to firm-level valuations. The economic validity of q^{AI}—its ability to predict investment and to decompose the average-marginal wedge—is assessed separately in the empirical analysis via investment regressions and event-level validations. Appendix E.7 provides the complete hyperparameter configuration, the grouped splitting procedure, and the evaluation algorithm.

3.6 Decomposing: Which Corpus is Most Informative About Marginal q

A distinctive feature of q^{AI} is that it is constructed from identifiable information sources, making it possible to ask which input groups covary most strongly with the resulting measure. Conventional q proxies—whether standard average Q or the intangible-adjusted Q^{tot} of Peters and Taylor (2017)—are opaque functions of market capitalization and book values, offering no window into what

information drives their variation. Because our pipeline assembles evidence from six distinct source groups, we can decompose the variation in q^{AI} into each source's marginal contribution.

An important caveat applies throughout: because q^{AI} is itself generated from these sources via the LLM, this decomposition describes how the pipeline weights its inputs, not a causal test of which sources improve investment prediction. The economic validation of q^{AI}—including its ability to predict investment beyond conventional proxies—is established separately in the investment regressions of Section 4. The decomposition here serves a different purpose: it provides transparency into the constructed measure and allows us to assess whether the pipeline's information weighting varies across industries in economically sensible ways.

Shapley-Owen R^2 decomposition. We implement this decomposition using the Shapley-Owen method, which allocates the R^2 of a regression among groups of regressors by averaging each group's marginal contribution across all possible orderings of entry (Huettner and Sunder, 2012). For each of the six source groups—the firm's own 10-K filings, same-industry peer 10-Ks, earnings conference calls, patent filings, Wall Street Journal articles, and Compustat financial variables—we construct a low-dimensional representation using the leading principal components of the source-specific text embeddings, retaining components that explain 80% of within-source variance. We then regress q^{AI} on the full set of source-group features and compute each group's Shapley-Owen share of the total R^2. The Shapley-Owen shares sum exactly to the model's R^2 and are invariant to the ordering in which source groups enter, ensuring that each source receives credit only for variation it contributes beyond what the other sources already explain.

[Insert Figure 3 Here.]

Full-sample results. Figure 3 reports the decomposition for the full sample. Text sources collectively account for more than 80% of the explained variation in q^{AI}. The firm's own 10-K filings are the single largest contributor (23.7%), followed by patent filings (18.7%), earnings conference calls (14.4%), peer firm 10-Ks (13.8%), and macroeconomic news from the Wall Street Journal (11.6%). Compustat financial variables—which include assets, sales, employment, leverage, profitability, and capital stock measures—explain only 17.8%. That every individual text source contributes comparably to or more than the entirety of Compustat suggests that q^{AI} derives economically

distinct information from textual evidence about strategic direction, innovation, competitive position, and the macroeconomic environment—content that is largely absent from standard financial variables.

[Insert Figure 4 Here.]

Cross-industry variation. Figure 4 reports the decomposition separately for five Fama-French industry groups, revealing systematic variation in the pipeline's information weighting. In high-tech and healthcare, patents contribute approximately 20% of the variation, consistent with innovation-driven investment opportunity sets in which the R&D pipeline is a first-order determinant of marginal project value. In manufacturing, contributions are most balanced across sources, spanning physical capacity, supply chains, and product markets. In services, peer firm 10-Ks are unusually important (~25%) while patents are negligible, and Compustat accounting variables contribute the least of any industry (~7%)—consistent with service-sector investment opportunities being shaped primarily by competitive dynamics rather than by tangible assets or patented technologies. These cross-industry patterns suggest that the pipeline's information weighting adapts to the structure of each industry's investment opportunity set, rather than applying a uniform weighting across all firms.

3.7 Validation: Similarity with Future Documents

The pipeline generates project proposals using only pre-period evidence. A natural validation question is whether these proposals bear any resemblance to what the firm actually does next. If q^{AI} captures genuine forward-looking investment opportunities, then firms with higher q^{AI} should subsequently pursue activities more closely aligned with the projects the system proposed. We test this by measuring the cosine similarity between generated project descriptions and the firm's subsequent 10-K filings and patent texts.

Measuring forward similarity. For each firm-year, we embed the generated project descriptions and the firm's future documents—10-K filings (Items 1 and 7) and patent texts—into the same vector space used for retrieval (Section 2.3) and compute cosine similarity. We standardize the resulting

scores to z-scores for cross-source comparability. To examine how forward similarity varies across the q^{AI} distribution, we partition firm-years into six tercile bins, split at $q^{AI} = 1$ for finer resolution near the break-even threshold, and report results at one-, two-, and three-year forward horizons. Because the retrieval stage enforces strict temporal filtering (Section 3.1), no future documents enter the evidence bundle, so any forward similarity cannot reflect direct textual overlap between retrieved and target texts. The validation is conducted against the firm's realized disclosures and patent filings—external benchmarks of actual corporate activity—rather than against the pipeline's own intermediate outputs.

A natural concern is that the test may partly capture textual persistence rather than forecasting: firms with stable strategies produce similar language across years, and the pipeline absorbs this persistence through the firm's own historical filings. While we cannot fully rule out this channel, we note that the test's power lies in the *cross-sectional gradient*—why higher-q^{AI} firms exhibit systematically greater forward similarity than lower-q^{AI} firms—rather than in the level of similarity for any individual firm. Moreover, the monotonic pattern across q^{AI} groups holds for patent filings, which reflect concrete R&D output rather than recycled managerial narrative, providing a harder test of the persistence explanation.

[Insert Figure 5 Here.]

Strategic disclosures. Figure 5 plots the standardized cosine similarity between generated project descriptions and the firm's future 10-K filings, by q^{AI} group. The pattern is monotonically increasing: firms in the highest q^{AI} group generate projects that are substantially more similar to their subsequent annual reports than firms in the lowest group. The pattern persists at two- and three-year forward horizons, though with some attenuation, suggesting that q^{AI} captures not only near-term investment plans but also the firm's medium-run strategic direction.

[Insert Figure 6 Here.]

Innovation output. Figure 6 reports the analogous test using the firm's future patent filings. The same monotonic pattern obtains: firms with higher q^{AI} subsequently file patents whose content is more similar to the generated project descriptions. Patent filings are particularly informative

because they reflect concrete R&D output—technologies, designs, and process innovations—rather than strategic narrative. The monotonic pattern in both dimensions—strategic disclosures and innovation output—is consistent with q^{AI} capturing forward-looking investment signals that extend across both the strategic and technological margins of the firm's opportunity set.

4 Empirical Results: Investment-q Relation

We now turn to the central empirical question: does q^{AI} predict corporate investment better than existing measures? We evaluate our AI-measured marginal q using the standard investment-q regression framework, comparing it against two benchmarks—standard average Q (Q^{std}, defined as the ratio of firm market value to the book value of physical capital) and intangible-adjusted total Q (Q^{tot}, following Peters and Taylor, 2017, which incorporates knowledge capital and organization capital into the replacement cost denominator). The analysis proceeds in two parts. At the firm level, we present baseline investment-q regressions, add cash flow controls, apply measurement error correction via the Erickson, Jiang, and Whited (2014) cumulant estimator, and report horse-race and ORIV robustness checks. We then turn to aggregate time-series evidence.

4.1 Firm-Level Investment-q Relation

Empirical setup. Following Peters and Taylor (2017), we estimate OLS panel regressions of the form

$$I_{it}/K_{i,t-1} = \beta \cdot q_{i,t-1} + \alpha_i + \eta_t + \epsilon_{it}, \tag{5}$$

where $I_{it}/K_{i,t-1}$ is the investment rate (in percent) for firm i in year t, $q_{i,t-1}$ is lagged Tobin's q, α_i are firm fixed effects, η_t are year fixed effects, and standard errors are clustered by firm. We measure five investment rates. Columns (1)–(4) of Table 3 scale investment by total capital (the sum of physical capital, measured as net PPE, and intangible capital, constructed via the perpetual-inventory method of Peters and Taylor, 2017): physical investment (capital expenditures), intangible investment (R&D plus a fraction of SG&A), total investment (the sum), and R&D investment. Column (5) uses the traditional investment rate (capital expenditures scaled by net PPE) to facilitate comparison with

earlier work that does not account for intangible capital.[4] Each row of the table reports a separate univariate regression of the investment rate on one of three q measures: Q^{std}, Q^{tot}, or q^{AI}. All q measures are lagged one year relative to the investment rate.

Baseline results. Table 3 presents the results. Three patterns emerge. First, the coefficient magnitudes increase sharply as we move from Q^{std} to Q^{tot} to q^{AI}. For total investment (column 3), the coefficient on Q^{std} is 0.012 and statistically insignificant—consistent with the longstanding finding that standard average Q is a weak predictor of investment in within-firm panel regressions (Erickson and Whited, 2000). (The one exception is the traditional CAPX/PPE specification in column 5, where Q^{std} is significant at the 5% level, likely reflecting the mechanical link between market-to-book and physical investment that is diluted when intangible investment is included.) The coefficient on Q^{tot} rises to 1.349 ($p < 0.01$), reflecting the gains from incorporating intangible capital documented in Peters and Taylor (2017). The coefficient on q^{AI} is 2.552 ($p < 0.01$), roughly twice that of Q^{tot}. This pattern holds across all five investment types.

Second, the within-firm R^2 comparisons confirm that q^{AI} captures more of the variation in investment rates. The bottom panel of Table 3 reports pairwise differences in within-firm R^2, with standard errors from a clustered bootstrap (1,000 replications, resampling firms).[5] The ΔR^2 between q^{AI} and Q^{std} ranges from 0.128 to 0.209, all significant at the 1% level. The ΔR^2 between Q^{tot} and Q^{std} is of comparable magnitude (0.105 to 0.168), consistent with the gains reported in Peters and Taylor (2017). The more demanding comparison is between q^{AI} and Q^{tot}: the ΔR^2 is 0.042 for total investment ($p < 0.01$), 0.030 for intangible investment ($p < 0.05$), and 0.023 for the standard investment rate ($p < 0.10$). While these incremental gains over Q^{tot} are modest in absolute terms, they are statistically significant and economically meaningful: the improvement reflects information about investment opportunities that is present in firm text but absent from the accounting aggregates underlying Q^{tot}.

Third, the coefficient hierarchy is itself informative about measurement quality. Under classical

[4] All investment rates and q measures are winsorized at the 1st and 99th percentiles. The sample comprises 10,616 firm-year observations for 455–458 firms over 1994–2022, restricted to firms with at least 25 years of data and excluding financials and utilities.

[5] The bootstrap resamples entire firm histories to preserve within-firm serial correlation. We test the null hypothesis that the within-firm R^2 values from two univariate regressions are equal; the reported standard errors are from the bootstrap distribution of the R^2 difference.

measurement error, OLS attenuates the true coefficient toward zero, with the attenuation factor equal to the signal-to-total variance ratio in the mismeasured regressor (Erickson and Whited, 2000, 2012). A q proxy with less measurement error should therefore produce a larger OLS coefficient. The fact that q^{AI} yields coefficients roughly twice as large as Q^{tot} provides an initial indication that our measure contains more signal about the true marginal return to capital—a hypothesis we test formally using the Erickson, Jiang, and Whited (2014) estimator below. What economic information drives this improvement? The Shapley-Owen decomposition reported in Section 3.6 shows that text sources collectively contribute more than 80% of q^{AI}'s variation, with the firm's own 10-K filings, patents, and peer disclosures each playing distinct roles. The improvement thus reflects forward-looking strategic and competitive information that backward-looking accounting data cannot encode.

[Insert Table 3 Here.]

Controlling for cash flow. A longstanding question in the investment literature is whether cash flow enters investment regressions because of genuine financing frictions (Fazzari, Hubbard, and Petersen, 1988) or because it proxies for investment opportunities that q measures with error (Kaplan and Zingales, 1997). Abel (2018) provide a sharp theoretical prediction: if q is well measured, cash flow carries no additional information about the firm's investment opportunity set, and its coefficient should be zero. Table 4 reports this test. Panel A reproduces the q^{AI} specification from Table 3. Panel B adds total cash flow (scaled by total capital) as a control. The coefficient on q^{AI} declines modestly—from 2.552 to 2.119 for total investment—but remains large and highly significant. Cash flow enters positively and significantly for physical, intangible, and total investment, but its coefficient (0.158 for total investment) is small relative to the q^{AI} coefficient. Notably, cash flow is insignificant for R&D investment (-0.002, $p > 0.10$), consistent with the view that R&D decisions respond primarily to long-run investment opportunities rather than contemporaneous cash availability.

These patterns are consistent with the prediction of Abel (2018): better measurement of q substantially reduces, though does not fully eliminate, the explanatory role of cash flow. The residual cash flow effect may reflect either remaining measurement imperfections in q^{AI} or genuine financing frictions—a distinction we do not attempt to resolve here but revisit in Section 5.

[Insert Table 4 Here.]

Measurement error correction. Even after the improvements documented above, OLS estimates remain attenuated by measurement error in all three q proxies. To quantify this attenuation and compare the signal content of q^{AI} and Q^{tot}, we apply the higher-order cumulant estimator of Erickson, Jiang, and Whited (2014). This estimator exploits the nonnormality of the mismeasured regressor—specifically, its third- and fourth-order cumulants—to identify the true regression coefficient without requiring external instruments, under the assumption that the measurement error is classical (mean-zero, independent of the true regressor and the outcome).[6] The estimator produces two diagnostic statistics: ρ^2, the signal-to-total variance ratio $(\text{var}(q^*)/\text{var}(q^{obs})$, where q^* denotes true marginal q and q^{obs} the observed proxy), and τ^2, the reliability of the proxy as measured by the slope coefficient in the reverse regression of q^* on q^{obs}. A higher ρ^2 indicates that a larger share of the proxy's variation reflects true investment opportunities; a lower τ^2 indicates that the proxy is less reliable as a linear predictor of the latent variable.

Table 5 reports the results. Panel A shows estimates using Q^{tot}; Panel B uses q^{AI}. Three findings emerge. First, the bias-corrected coefficients are substantially larger than their OLS counterparts. For total investment, the corrected coefficient on q^{AI} is 4.854, compared to the OLS estimate of 2.552—implying that OLS recovers only about 53% of the true investment-q elasticity, with the remaining 47% lost to attenuation. This attenuation is even more severe for Q^{tot}, where the corrected coefficient (2.955) is more than double the OLS estimate (1.349). The magnitudes confirm the finding of Erickson and Whited (2000) and Erickson and Whited (2012) that measurement error in q produces economically large bias.

Second, q^{AI} exhibits higher signal content than Q^{tot}. The ρ^2 for q^{AI} ranges from 0.271 (physical investment) to 0.441 (standard investment rate), while ρ^2 for Q^{tot} ranges from 0.195 to 0.318. Panel C reports the pairwise difference: $\Delta\rho^2$ (AI − tot) is positive across all five specifications, and statistically significant for intangible, total, R&D, and standard investment. This comparison rests on the maintained assumption that both proxies are noisy measures of the same latent object—the

[6] We implement the estimator using the `xtewreg` command in Stata with fourth-order cumulants and conditional maximum likelihood, following the recommendations in Erickson, Jiang, and Whited (2014). The specification includes firm fixed effects but not year fixed effects, as the cumulant estimator requires a balanced panel structure within each fixed-effects group. The OLS within-R^2 values reported in the bootstrap rows of Table 5 are from firm-FE-only regressions for comparability.

firm's true marginal return to capital. Under this assumption, the higher ρ^2 for q^{AI} confirms that our text-based measure captures more of the variation in the true marginal q, not merely that it is internally consistent.[7]

Third, the $\Delta\tau^2$ differences between q^{AI} and Q^{tot} are small and mostly insignificant. The overall pattern suggests that the improvement in q^{AI} comes primarily from capturing more signal about true investment opportunities (ρ^2), rather than from reducing the absolute noise level in the proxy.

We note two caveats. First, the cumulant estimator assumes classical measurement error—that the noise in the q proxy is mean-zero and uncorrelated with both true q and the regression error. For Q^{tot}, this is a standard and well-studied assumption (Erickson and Whited, 2012). For q^{AI}, which is a generated measure, systematic biases in the LLM's project valuations could violate this assumption. The validation exercises in Section 3—showing that q^{AI} is predominantly driven by retrieved text rather than parametric model knowledge, and that the measure is robust to firm-identity anonymization—provide some reassurance, but we cannot fully rule out non-classical error. Second, the specifications in Table 5 include firm fixed effects but not year fixed effects, as required by the estimator's panel structure. The OLS results in Tables 3 and 4, which include both firm and year fixed effects, serve as the primary specification.

[Insert Table 5 Here.]

Horse race and ORIV estimates. We conduct two additional exercises, reported in the appendix, to further assess the relative information content of q^{AI} and Q^{tot}. Appendix Table A2 enters both measures simultaneously. In the joint specification for total investment, the coefficient on q^{AI} declines modestly (from 2.552 to 2.121) but remains highly significant, while the coefficient on Q^{tot} drops by roughly three-quarters (from 1.349 to 0.340), falling from the 1% to the 5% significance level in three of five specifications. The joint R^2 (0.650) is virtually identical to the q^{AI}-only R^2 (0.649), indicating that Q^{tot} contributes negligible incremental explanatory power.

Appendix Table A3 applies the Obviously Related Instrumental Variables (ORIV) estimator of Gillen, Snowberg, and Yariv (2019), which treats q^{AI} and Q^{tot} as two noisy proxies for the same

[7] If q^{AI} and Q^{tot} proxy for different latent constructs, the ρ^2 comparison would not be well-defined. Our maintained assumption is that both are imperfect proxies for the same theoretical object—the shadow value of capital that drives investment decisions. The economic validity of this assumption is supported by the fact that both measures predict the same investment outcomes and that the bias-corrected coefficients represent the same structural parameter (the investment-q elasticity) under the null.

latent variable and uses each as an instrument for the other. The ORIV-corrected coefficients (e.g., 2.486 for total investment) are close to the q^{AI} OLS estimates and substantially above the Q^{tot} OLS estimates, consistent with q^{AI} being the less-attenuated proxy. Both exercises reinforce the cumulant-estimator finding: q^{AI} captures more of the variation in true marginal q than Q^{tot}.

4.2 Aggregate Investment-q Relation

Empirical setup. We complement the firm-level panel evidence with aggregate time-series analysis, following Philippon (2009) and Andrei, Mann, and Moyen (2019). We construct value-weighted aggregate series by collapsing the firm-level panel, weighting by lagged firm market value, and regress the aggregate investment rate on lagged aggregate q.[8]

Results. Figure 2 plots the value-weighted aggregate total investment rate alongside aggregate q^{AI} over the sample period. The two series co-move closely, with a time-series correlation of 0.63 and a simple regression R^2 of 0.40. Both peak around the dot-com boom (~2000) and decline through the 2000s. Post-2010, however, the series partially diverge: q^{AI} recovers while the aggregate investment rate continues to drift downward.

This partial divergence is economically informative rather than a failure of the measure. It is important to distinguish two phenomena. The "investmentless growth" puzzle in its original formulation refers to the divergence between *average* Q and investment (Gutiérrez and Philippon, 2017)—a gap that our wedge decomposition in Section 5 attributes primarily to capitalized rents and omitted intangibles inflating average Q beyond what investment opportunities justify. By contrast, the milder post-2010 divergence between q^{AI} and aggregate investment likely reflects a narrower channel: the shift toward intangible-intensive production means that true capital formation increasingly takes forms—organizational investment, data assets, software development—that are imperfectly captured even by our broad investment measure, which includes R&D but not all categories of intangible spending (De Ridder, 2024). In other words, the remaining gap between q^{AI} and measured investment points to unmeasured investment rather than to mismeasured q.

[8] Our sample of 455–458 firms with at least 25 years of data is weighted toward large, established firms that account for a substantial share of aggregate market capitalization and capital formation. While this selection ensures reliable firm-level estimation, it means the aggregate series may not capture the full cross-section of investment activity, particularly among young or small firms.

The aggregate results also connect to the finding of Andrei, Mann, and Moyen (2019) that the investment-Q relationship was weak before the mid-1990s but tightened substantially thereafter, driven by the rise of innovation-intensive firms. Our firm-level results are consistent with this pattern: q^{AI} draws heavily on text sources—patents, peer disclosures, and forward-looking strategic narratives—that are most informative for firms whose investment opportunities are poorly captured by accounting data alone.

[Insert Figure 2 Here.]

5 New Economics of the Marginal Project

Our method identifies the marginal project itself—not just its value. Unlike scalar q measures, q^{AI} comes with a textual project description, enabling us to classify marginal projects by type, measure their compatibility with existing operations, and distinguish organic expansion from acquisition-driven growth. We exploit these features to test three refined predictions of investment theory: *what* the investment is (heterogeneity by type), *how far* it departs from existing operations (intensive versus extensive margin), and *how* it is implemented (build versus buy).[9]

We then turn to a different question: if q^{AI} outperforms Q in predicting investment, why does Q fail? Having a direct measure of marginal q allows us to conduct a residual analysis of the wedge $Q - q^{AI}$, identifying which firm characteristics—capitalized rents, omitted intangibles, and financial frictions—are statistically associated with the gap between average and marginal valuations.

5.1 Heterogeneity: Investment Type

If q^{AI} captures economically meaningful, project-specific information—rather than a generic investment signal correlated with all forms of capital formation—then the investment-q elasticity should vary with the match between the type of marginal project the LLM identifies and the type of investment the firm undertakes. When the system proposes a physical expansion (e.g., a new

[9]Because q^{AI} is constructed from an LLM pipeline rather than observed directly, standard OLS standard errors may understate true estimation uncertainty. We address this concern in the Online Appendix by implementing bootstrap procedures that resample the generation process and by examining sensitivity to model temperature and prompt specification. The qualitative findings reported in this section are robust to these adjustments.

production facility), physical investment should be more responsive to q^{AI}—and likewise, R&D investment should respond more strongly when the system identifies an innovation-oriented project (e.g., a new drug candidate). This test directly validates whether q^{AI} carries project-level content absent from any scalar q measure.

Empirical setup. We classify LLM-generated project descriptions into three categories—physical capital, intangible capital, and R&D—using keyword dictionaries applied to the text of the three generated projects for each firm-year. The dictionaries, reported in Appendix Table A1, cover standard terms for plant and equipment (physical), software and organizational assets (intangible), and research activities (R&D); a project may be assigned to multiple categories.[10] We estimate:

$$I^j_{it}/K_{i,t-1} = \beta_0 \cdot q^{AI}_{i,t-1} + \beta_1 \cdot \mathbf{1}(\text{Type } j)_{i,t-1} + \beta_2 \cdot q^{AI}_{i,t-1} \cdot \mathbf{1}(\text{Type } j)_{i,t-1} + \alpha_i + \eta_t + \epsilon_{it}, \qquad (6)$$

where $I^j_{it}/K_{i,t-1}$ is the type-j investment rate (physical, intangible, or R&D investment scaled by total capital), and $\mathbf{1}(\text{Type } j)$ is an indicator equal to one if the LLM-generated marginal project is classified as type j. A positive β_2 indicates that the type-specific investment rate responds more to q^{AI} when the marginal project matches that investment category.[11]

Results. Table 6 reports the estimates. The interaction coefficient β_2 is positive and statistically significant across all three investment types: 0.132 ($p < 0.01$) for physical investment, 0.145 ($p < 0.05$) for intangible investment, and 0.228 ($p < 0.01$) for R&D. These magnitudes are economically meaningful. When the LLM identifies an R&D-oriented marginal project, the R&D investment-q elasticity rises by 26% (from 0.883 to 1.111); when it identifies a physical project, the physical investment-q elasticity rises by 15% (from 0.854 to 0.986). The R&D result is particularly striking: R&D investment decisions appear especially sensitive to whether the marginal project involves innovation, consistent with the view that R&D spending is more discretionary and project-specific than routine capital expenditure.

[10] We validate the keyword classification by hand-coding a random sample of 200 LLM-generated project descriptions. The agreement rate between dictionary-based and manual classifications exceeds 85% across all three categories. Results are reported in the Online Appendix.

[11] Following the neoclassical q-theory prediction that investment should depend only on q, our specifications omit cash flow. Adding lagged cash flow as a control does not materially affect the interaction coefficients reported in Tables 6–7; see the Online Appendix.

These results confirm that q^{AI} is not merely a better-measured version of Q—it contains qualitatively distinct information about the *composition* of investment opportunities. While one could in principle combine traditional q measures with external project-type indicators (e.g., capex versus R&D splits from financial statements), no existing scalar q measure endogenously identifies the nature of the marginal project at the granularity our method achieves.

[Insert Table 6 Here.]

5.2 Heterogeneity: Compatibility with Existing Operations

Jovanovic and Rousseau (2014) argue that the investment-q relation is state-dependent: the marginal response of investment to q depends on the compatibility of the new project with the firm's existing capital stock. When the marginal project is an *intensive*-margin expansion—scaling up activities closely related to the firm's current operations—the standard prediction holds: higher q elicits more investment, as the new capital complements what is already in place. When the marginal project instead requires entry into new strategic territory (the *extensive* margin), the investment-q elasticity compresses. Adopting a project that is incompatible with existing operations may devalue installed capital, require costly organizational restructuring, or entail implementation frictions that attenuate the marginal response.

This framework incorporates cannibalization into an otherwise standard q theory: the cost of the marginal project depends not only on the project itself but also on its relation with the firm's existing operations. Jovanovic and Rousseau (2014) provide aggregate evidence using IPO-based measures of extensive investment, but the prediction has not been tested at the firm level—in part because no existing q measure characterizes the marginal project's compatibility with the firm's current activities. Our text-based approach enables exactly this test. Because we observe the generated project description, we can measure its similarity to the firm's existing operations and directly test whether the investment-q elasticity varies along the intensive-extensive margin.

Empirical setup. We measure compatibility as the cosine similarity between the embedding of the LLM-generated project description and the embeddings of the firm's contemporaneous 10-K filings (Items 1 and 7A) and earnings conference call transcripts. High similarity indicates that the

marginal project falls within the firm's established strategic domain (intensive margin); low similarity indicates entry into new territory (extensive margin). We define $\mathbf{1}(\text{High Compatibility})_{i,t-1}$ as an indicator equal to one if the cosine similarity exceeds the sample median, and estimate:

$$I_{it}/K_{i,t-1} = \beta_0 \cdot q^{AI}_{i,t-1} + \beta_2 \cdot q^{AI}_{i,t-1} \cdot \mathbf{1}(\text{High Compatibility})_{i,t-1} + \alpha_i + \eta_t + \epsilon_{it}. \tag{7}$$

Results. Table 7 reports the estimates. The baseline coefficient β_0—the investment-q elasticity for low-compatibility (extensive-margin) projects—is 1.882 ($p < 0.01$) for total investment. The interaction term β_2 is 0.679 ($p < 0.01$), implying that the total investment-q elasticity is 36% larger when the marginal project is compatible with existing operations (2.561 versus 1.882). The pattern is consistent across investment types: the interaction is 0.488 ($p < 0.01$) for intangible investment, 0.471 ($p < 0.01$) for R&D, and 0.547 ($p < 0.05$) for the standard investment rate. Physical investment is the exception: the interaction (0.124) is positive but statistically insignificant.

The asymmetry across investment types is economically informative. Physical capital—factories, equipment, warehouses—is relatively fungible across projects; a manufacturing line can often be repurposed without major restructuring. In contrast, intangible and R&D investments are inherently project-specific: knowledge capital, organizational processes, and innovation pipelines built for one strategic direction are costly to redirect toward an incompatible domain. The stronger compatibility effect for intangible and R&D investment is therefore consistent with the mechanism in Jovanovic and Rousseau (2014): reallocation frictions are most binding for capital types that are tightly coupled to the firm's existing strategic position.

We note two caveats. First, compatibility as measured by textual similarity could partly reflect retrieval quality rather than true economic compatibility: firms with richer text data may produce higher-similarity project descriptions simply because the retrieval system has more material to draw on. Firm fixed effects absorb time-invariant differences in text availability, but within-firm variation in text richness could remain a confound. The differential pattern across investment types—strong effects for intangible and R&D but not physical investment—mitigates this concern, since retrieval quality should affect all investment types similarly.

Second, low compatibility could signal genuinely low-return projects rather than reallocation frictions per se: a project that is distant from existing operations may have low expected returns

independent of transition costs. Disentangling friction-based from return-based explanations for the extensive-margin discount is beyond the scope of this paper, but the pattern is consistent with the friction channel emphasized by Jovanovic and Rousseau (2014). A natural follow-up test—examining whether low-compatibility projects earn lower realized returns conditional on q^{AI}—would directly validate the reallocation-friction mechanism versus explanations based on unobserved project quality. We leave this to future work.

[Insert Table 7 Here.]

5.3 Heterogeneity: Build versus Buy

The q-theory of mergers (Jovanovic and Rousseau, 2002, 2003) predicts that high-q firms acquire low-q targets when buying existing assets is cheaper than building new capital. This build-versus-buy dimension is distinct from the intensive-extensive margin tested above: acquisitions represent a specific implementation channel for growth that involves its own frictions—integration costs, regulatory approval, cultural mismatch—beyond the adjustment costs of organic investment. By classifying our LLM-generated projects as "build" (organic development, construction, R&D) or "buy" (acquisitions, mergers, licensing-in), we can test whether the composition of the marginal project maps onto actual corporate behavior and whether the investment-q relation differs across these channels.

Empirical setup. We classify generated project descriptions as "buy" or "build" using keyword dictionaries (Appendix Table A1, Panel D). Buy keywords include terms related to acquisitions, mergers, and purchasing; build keywords cover organic development, construction, and commercialization. We then run four specifications. Column (1) tests whether the LLM's "buy" classification predicts actual M&A activity by regressing an indicator for whether the firm undertakes an acquisition in year t on $\mathbf{1}(\text{Buy})_{i,t-1}$, with firm and year fixed effects. Columns (2)–(4) examine the investment-q relation separately for the M&A and non-M&A subsamples.

Results. Table 8 reports the results.

Column (1) provides a reduced-form validation of the project classification: when the LLM identifies a buy-type marginal project, the probability that the firm actually undertakes an acquisition

in the following year increases by 4.7 percentage points ($p < 0.10$). Given a baseline M&A frequency of 46.3%, this represents a 10% increase in the likelihood of acquisition. While modest in magnitude and only marginally significant, this finding confirms that the generated project descriptions carry information about subsequent corporate actions beyond what firm and year fixed effects absorb.

Columns (2) and (3) restrict the sample to firm-years with at least one acquisition. Among these acquirers, q^{AI} predicts both M&A-specific investment (coefficient 2.340, $p < 0.01$) and total investment (3.475, $p < 0.01$). The total investment coefficient in this subsample exceeds the full-sample estimate (2.552 from Table 3), consistent with the prediction of Jovanovic and Rousseau (2002) that M&A investment is more q-elastic than organic capital formation.[12]

Column (4) restricts the sample to firm-years with no M&A activity. The coefficient on q^{AI} remains large and highly significant (2.459, $p < 0.01$), confirming that our measure captures organic investment opportunities as well. The ability to predict investment in both acquirer and non-acquirer subsamples rules out the concern that q^{AI} is mechanically driven by M&A valuation effects.

Taken together with the compatibility results in Section 5.2, these findings paint a richer picture of how implementation channels shape the investment-q relation. Section 5.2 showed that extensive-margin projects—those incompatible with existing operations—exhibit a lower investment-q elasticity. Here, we find that the LLM's buy classification predicts actual M&A and that the investment-q relation is strong in both acquirer and non-acquirer subsamples, confirming that q^{AI} captures genuine investment opportunities across implementation modes. The higher coefficient among acquirers (3.475 versus 2.459) is consistent with M&A providing a high-powered but selective channel for acting on investment opportunities—firms that acquire may be those best positioned to exploit high-q projects through external growth. This connects to the recent finding of Grullon, Ikenberry, and Verdi (2025) that anticipated underutilization of new capital drives a wedge between average Q and the marginal return to investment. Our project-level decomposition complements their insight: we can directly observe *what kind of project* drives the marginal investment decision

[12] We note that the sample split on realized M&A status conditions on an endogenous outcome. The larger coefficient among acquirers could reflect either genuine q-elasticity differences or selection: firms that undertake M&A may differ along unobserved dimensions that make them more q-responsive. The comparison across subsamples should therefore be interpreted as descriptive rather than causal.

and how its characteristics shape the investment response.

[Insert Table 8 Here.]

5.4 Dissecting the Average Q–Marginal q^{AI} Wedge

The results above establish that q^{AI} outperforms average Q in predicting investment and that the investment-q elasticity varies with project characteristics. A natural follow-up is: what does average Q capture that marginal q does not? The wedge $Q - q^{AI}$ isolates components of the firm's market-to-capital ratio that are orthogonal to the valuation of marginal investment opportunities—the divergence between the firm's average valuation ratio and the value implied by its marginal projects. Theory identifies three economically distinct sources.

First, capitalized market-power rents: firms with pricing power earn rents that inflate market value beyond what investment opportunities alone justify (Lindenberg and Ross, 1981; De Loecker, Eeckhout, and Unger, 2020). These rents are capitalized in the stock price and thus in average Q, but they do not raise the return to the marginal project—indeed, they may reduce the incentive to invest if the firm can earn supernormal profits on its existing capital stock. Second, omitted intangible capital: firms whose productive capital consists substantially of knowledge, brand equity, or organizational assets have market values that reflect capital not captured in the replacement cost denominator of standard Q (Peters and Taylor, 2017; Ewens, Peters, and Wang, 2024). The shift toward intangible-intensive production has widened this gap over time (De Ridder, 2024). Third, scale effects and financial frictions: firms facing decreasing returns to scale or binding financial constraints have market values that embed option values, franchise values, or growth expectations that diverge from the marginal project's value.

Crouzet and Eberly (2023) provide the theoretical foundation for decomposing the investment gap. Their Q+ framework identifies three channels: flow rents to physical capital, omitted intangible capital stocks, and an interaction term capturing rents accruing specifically to intangible capital. They document that the interaction term accounts for a substantial and growing share of the aggregate gap, particularly in technology-intensive sectors. Our contribution is to provide among the first firm-level empirical decompositions of this wedge, using a direct measure of marginal q rather than a structural estimate derived from accounting aggregates. We focus on rents, intangibles,

and financial frictions as the primary channels in our firm-level analysis, and we emphasize that our regression-based approach is a *residual analysis*—it identifies which firm characteristics are statistically associated with the wedge—rather than a structural decomposition in the sense of Crouzet and Eberly (2023).

Empirical setup. We regress the wedge $Q^{tot}_{i,t-1} - q^{AI}_{i,t-1}$ on firm characteristics that proxy for each theoretical channel:

$$(Q^{tot}_{i,t-1} - q^{AI}_{i,t-1}) = \gamma_1 \cdot \text{Profit Margin}_{it} + \gamma_2 \cdot \text{Intangible Intensity}_{it} + \gamma_3 \cdot \log K^{tot}_{it} + \gamma_4' \mathbf{X}_{it} + \alpha_i + \eta_t + \epsilon_{it}, \quad (8)$$

where Profit Margin proxies for capitalized rents (higher profitability inflates average Q beyond what investment opportunities justify), Intangible Intensity (the share of intangible investment in total investment) captures omitted intangible capital, $\log K^{tot}$ captures scale-related effects and decreasing returns, and $\mathbf{X}$ includes the Whited and Wu (2006) WW index and book leverage as measures of financial frictions.

Results. Table 9 reports the results. Columns (1)–(4) enter each channel separately; column (5) includes all regressors jointly.

Profit margin is strongly and positively associated with the wedge: the coefficient is 0.064 ($p < 0.01$) in column (1) and 0.053 ($p < 0.01$) in the joint specification, consistent with the prediction that capitalized rents inflate average Q without raising the marginal return to investment.

Intangible intensity is the single largest driver of the wedge: the coefficient is 0.126 ($p < 0.01$) in column (2) and 0.127 ($p < 0.01$) jointly. Firms with a greater share of intangible investment have larger gaps between average Q and marginal q^{AI}, reflecting the omission of intangible capital from the replacement cost base used to construct Q. This finding complements Peters and Taylor (2017), who show that accounting for intangible capital improves the investment-q relation from the Q side. Our result documents the same phenomenon from the other direction: the portion of firm value attributable to unmeasured intangibles shows up as a wedge between Q and our text-based marginal q.

The financial constraints measures contribute in an economically interpretable direction. The WW index enters negatively (-1.543, $p < 0.01$ in column 4; -2.282, $p < 0.01$ jointly): more

financially constrained firms have *smaller* wedges. This pattern is consistent with constraints compressing the growth options and franchise value that inflate average Q above marginal project values. Because q^{AI} is constructed from project-level text—reflecting the value of the firm's most plausible investment opportunities, not what the market prices into its equity—it captures the quality of available projects rather than the firm's ability to finance them. A financially constrained firm may face attractive marginal projects that it cannot fully fund; q^{AI} reflects this opportunity, while average Q is depressed by the market's assessment that financing frictions will prevent full exploitation. Constraints thus narrow the wedge primarily by depressing Q rather than by affecting q^{AI}. In the joint specification, log total capital becomes significantly negative (-1.310, $p < 0.01$), suggesting that conditional on rents and intangible intensity, larger firms have smaller wedges—a pattern consistent with decreasing returns to scale compressing average Q toward marginal q as the capital stock grows.

[Insert Table 9 Here.]

Figure 7 plots the decomposition over time and across industries. In the aggregate time series (panel a), intangible intensity is the largest component throughout and grows steadily after 2010, while the contribution of rents drives the dot-com peak around 2000 and the post-2015 surge. The cross-industry panels reveal systematic variation that echoes the findings of Crouzet and Eberly (2023). In high-tech (panel c), intangible omission dominates, consistent with these firms' heavy reliance on unmeasured knowledge and organizational capital. In manufacturing (panel b), contributions are more balanced across channels. In services (panel e), intangible intensity is particularly pronounced, while in healthcare (panel f), both rents and intangibles contribute substantially—consistent with the pharmaceutical industry's combination of high market power and large unmeasured R&D capital.

[Insert Figure 7 Here.]

These results reveal that the two improvements pursued in prior work—econometric correction for measurement error (Erickson and Whited, 2000, 2012) and adjustment for intangible capital (Peters and Taylor, 2017)—address only two of the three sources of the average-marginal wedge. Measurement error correction can recover the true coefficient under classical noise, and intangible

adjustment narrows the gap between the Q denominator and the true capital stock. Neither, however, addresses capitalized rents, which inflate the Q numerator with components of firm value that do not reflect investment opportunities. Our approach sidesteps this wedge: by constructing marginal q from project-level text rather than accounting aggregates, the resulting measure is orthogonal to the pricing of rents, intangibles, or financial frictions embedded in Q. The limited predictive power of Q relative to q^{AI} is thus not a failure of q theory but a measurement problem with identifiable, quantifiable sources.

6 Conclusion

For fifty years, empirical work on investment has substituted average Q for marginal q and found it wanting. These average Q approaches take a top-down view and construct Q using aggregate accounting data without going into firm's investment opportunities. This paper reverses the process and revisits the investment-q relation. We construct marginal q from the bottom up: for each firm-year, a retrieval-augmented generation system assembles textual evidence from filings, patents, earnings calls, peer disclosures, and news, and a large language model proposes the firm's most plausible next projects along with their estimated values and costs. Text carries more than 80 percent of the resulting measure's variation. Accounting aggregates alone carry less than 18 percent.

q^{AI} substantially outperforms average Q and intangible-adjusted total Q in predicting investment, both at the firm level and in the aggregate. When q^{AI} replaces conventional measures, cash flow largely loses its explanatory power—consistent with Abel (2018)'s prediction that a well-measured q should be sufficient. The weak empirical performance of q theory reflects how we have measured q, not a failure of the theory itself.

Observing the marginal project—not just its shadow price—opens tests that scalar q measures cannot support. The investment-q elasticity rises when the type of project the model identifies matches the type of investment the firm undertakes. It is larger for projects compatible with the firm's existing operations than for extensive-margin expansion, as the adjustment-cost framework of Jovanovic and Rousseau (2014) predicts. Decomposing the wedge between average Q and q^{AI}, we trace the gap to three distinct sources: capitalized rents, omitted intangible capital, and financial frictions. The two fixes proposed in prior work—econometric correction for measurement error

and inclusion of intangibles in the capital stock—address only two of these. Rents, which inflate the numerator of Q without raising the return to the marginal project, remain untouched.

The methodology has implications beyond q. The core of the approach is a retrieval system that selects, for each observation, the particular combination of texts most relevant to the economic quantity of interest. Which texts matter varies: 10-K filings dominate for manufacturing firms, patents for technology firms, and conference calls for service firms. This context-dependence is not an artifact; it is the reason the pipeline outperforms fixed-corpus alternatives. Any latent economic object that leaves dispersed, heterogeneous traces in firm text is a candidate for the same architecture.

Many of the quantities that empirical finance treats as latent fit this description. Financial constraints are discussed in earnings calls and covenant disclosures. Agency problems surface in proxy statements. Risk exposures are narrated in 10-K risk factors. In each case, the field relies on accounting-based proxies whose limitations are well documented. The contribution of this paper is to show, in one important domain, that moving measurement from accounting aggregates to the text that firms produce can close a large part of the gap between what theory predicts and what empirics deliver.

Variable Definition

Variable	Definition and Construction
Panel A: Investment Rates	
$I_{it}^{phys}/K_{i,t-1}^{tot}$	Physical investment rate (%). Capital expenditures (Compustat: `capx`) divided by lagged total capital, multiplied by 100.
$I_{it}^{int}/K_{i,t-1}^{tot}$	Intangible investment rate (%). Intangible investment (R&D plus a fraction of SG&A, following Peters and Taylor, 2017) divided by lagged total capital, multiplied by 100.
$I_{it}^{tot}/K_{i,t-1}^{tot}$	Total investment rate (%). Sum of physical and intangible investment divided by lagged total capital, multiplied by 100.
$I_{it}^{R\&D}/K_{i,t-1}^{tot}$	R&D investment rate (%). R&D expenditure (`xrd`) divided by lagged total capital, multiplied by 100.
$I_{it}^{phys}/K_{i,t-1}^{phys}$	Standard investment rate (%). Capital expenditures divided by lagged net PPE (`ppent`), multiplied by 100.
$I_{it}^{M\&A}/K_{i,t-1}^{tot}$	M&A investment rate (%). Acquisitions (`aqc`) divided by lagged total capital, multiplied by 100.
$\mathbf{1}(\text{M\&A})_{it}$	Indicator equal to one if the firm undertakes an acquisition (`aqc` > 0) in year t.
Panel B: q Measures and Controls	
Q^{std}	Standard average Q. Market value of the firm divided by the book value of physical capital (net PPE). Market value is equity market capitalization plus book value of liabilities.
Q^{tot}	Intangible-adjusted total Q, following Peters and Taylor (2017). Market value of the firm divided by total capital (physical plus intangible). Intangible capital includes knowledge capital (perpetual inventory of R&D) and organization capital (perpetual inventory of a fraction of SG&A).
q^{AI}	AI-measured marginal q. For each firm-year, three project-level ratios of estimated market value to implementation cost are generated by the LLM-RAG pipeline (Section 3); q^{AI} is their simple average.
Total Cash Flow	Total cash flow scaled by total capital (%). Computed as (net income + depreciation + intangible investment $\times$ 0.7) divided by lagged total capital, multiplied by 100.
Panel C: Capital Stock Components	
K^{phys}	Physical capital. Net property, plant, and equipment (`ppent`).
K^{int}	Intangible capital. Sum of knowledge capital (perpetual inventory of R&D, 15% depreciation rate) and organization capital (perpetual inventory of 30% of SG&A, 15% depreciation rate), following Peters and Taylor (2017).
K^{tot}	Total capital. $K^{phys} + K^{int}$.
V	Market value of the firm. Equity market capitalization plus book value of total liabilities.
Panel D: Heterogeneity and Wedge Decomposition Variables	
$\mathbf{1}$(Project Type j)	Indicator equal to one if the LLM-generated marginal project is classified as type $j \in$ {physical, intangible, R&D} based on keyword dictionaries applied to the project description (Appendix Table A1).
$\mathbf{1}$(High Compat.)	Indicator equal to one if the cosine similarity between the LLM-generated project description and the firm's contemporaneous earnings conference call transcripts exceeds the sample median.
$\mathbf{1}$(Buy)	Indicator equal to one if the LLM-generated project is classified as acquisition-based using keyword dictionaries (Appendix Table A1, Panel D).

Continued on next page

Variable Definition (continued)

Variable	**Definition and Construction**
Profit Margin	Gross profit margin (%). (Revenue − COGS) / Revenue, floored at zero, multiplied by 100.
Intangible Intensity	Share of intangible capital in total capital (%). $K^{int}/(K^{int} + K^{phys})$, multiplied by 100.
log K^{tot}	Natural logarithm of total capital.
WW Index	Whited and Wu (2006) financial constraints index, standardized to zero mean and unit variance.
Book Leverage	Ratio of book value of debt to book value of total assets.

References

Abel, A. B. 2018. The effects of q and cash flow on investment in the presence of measurement error. *Journal of Financial Economics* 128:363–77.

Abel, A. B., and O. J. Blanchard. 1986. A dynamic model of investment and capacity utilization. *The Quarterly Journal of Economics* 101:429–44.

Andrei, D., W. Mann, and N. Moyen. 2019. Why did the q theory of investment start working? *Journal of Financial Economics* 133:251–72.

Baker, S. R., N. Bloom, and S. J. Davis. 2016. Measuring economic policy uncertainty. *The Quarterly Journal of Economics* 131:1593–636.

Belo, F., V. D. Gala, J. Salomao, and M. A. Vitorino. 2022. Decomposing firm value. *Journal of Financial Economics* 143:619–39.

Bybee, L., B. T. Kelly, and Y. Su. 2023. Narrative asset pricing: Interpretable systematic risk factors from news text. *The Review of Financial Studies* 36:4759–87.

Caballero, R. J. 1999. Aggregate investment. *Handbook of Macroeconomics* 1:813–62.

Crouzet, N., and J. C. Eberly. 2023. Rents and intangible capital: A q+ framework. *Journal of Finance* 78:507–45.

De Loecker, J., J. Eeckhout, and G. Unger. 2020. The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics* 135:561–644.

De Ridder, M. 2024. Market power and innovation in the intangible economy. *American Economic Review* 114:199–251.

Erickson, T., C. H. Jiang, and T. M. Whited. 2014. Minimum distance estimation of the errors-in-variables model using linear cumulant equations. *Journal of Econometrics* 183:211–21.

Erickson, T., and T. M. Whited. 2000. Measurement error and the relationship between investment and q. *Journal of Political Economy* 108:1027–57.

———. 2012. Treating measurement error in Tobin's q. *Review of Financial Studies* 25:1286–329.

Ewens, M., R. H. Peters, and S. Wang. 2024. Measuring intangible capital with market prices. *Management Science* 71:407–27.

Fazzari, S. M., R. G. Hubbard, and B. C. Petersen. 1988. Financing constraints and corporate investment. *Brookings Papers on Economic Activity* 1988:141–206.

Gabaix, X., R. S. Koijen, R. Richmond, and M. Yogo. 2024. Asset embeddings. *Available at SSRN 4507511* .

Gala, V., J. F. Gomes, and T. Liu. 2025. Marginal q. *Journal of Finance* Forthcoming.

Gentzkow, M., B. Kelly, and M. Taddy. 2019. Text as data. *Journal of Economic Literature* 57:535–74.

Gillen, B., E. Snowberg, and L. Yariv. 2019. Experimenting with measurement error: Techniques with applications to the Caltech cohort study. *Journal of Political Economy* 127:1826–63.

Grullon, G., D. L. Ikenberry, and R. S. Verdi. 2025. Excess capacity, marginal q, and corporate investment. *The Journal of Finance* 80:921–61.

Gutiérrez, G., and T. Philippon. 2017. Investmentless growth: An empirical investigation. *Brookings Papers on Economic Activity* 2017:89–169.

Hassan, T. A., S. Hollander, L. van Lent, and A. Tahoun. 2019. Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics* 134:2135–202.

Hayashi, F. 1982. Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica* 50:213–24.

Hoberg, G., and G. Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124:1423–65.

Hubbard, R. G. 1998. Capital-market imperfections and investment. *Journal of Economic Literature* 36:193–225.

Huettner, F., and M. Sunder. 2012. Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values. *Economics Letters* 117:107–9.

Jha, M., J. Qian, M. Weber, and B. Yang. 2024. Generative ai, managerial expectations, and economic activity. *arXiv preprint arXiv:2410.03897* .

Jovanovic, B., and P. L. Rousseau. 2002. The q-theory of mergers. *American Economic Review* 92:198–204.

———. 2003. The q-theory of ipos .

———. 2014. Extensive and intensive investment over the business cycle. *Journal of Political economy* 122:863–908.

Kaplan, S. N., and L. Zingales. 1997. Do investment-cash flow sensitivities provide useful measures of financing constraints? *The Quarterly Journal of Economics* 112:169–215.

Kim, A. G., M. Muhn, and V. V. Nikolaev. 2024. Financial statement analysis with large language models. *Working Paper* University of Chicago Booth School of Business.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman. 2017. Technological innovation, resource allocation, and growth. *The quarterly journal of economics* 132:665–712.

Lindenberg, E. B., and S. A. Ross. 1981. Tobin's q ratio and industrial organization. *Journal of Business* 54:1–32.

Loughran, T., and B. McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance* 66:35–65.

Peters, R. H., and L. A. Taylor. 2017. Intangible capital and the investment-q relation. *Journal of Financial Economics* 123:251–72.

Philippon, T. 2009. The bond market's q. *The Quarterly Journal of Economics* 124:1011–56.

Sarkar, S. K. 2025. Economic representations.

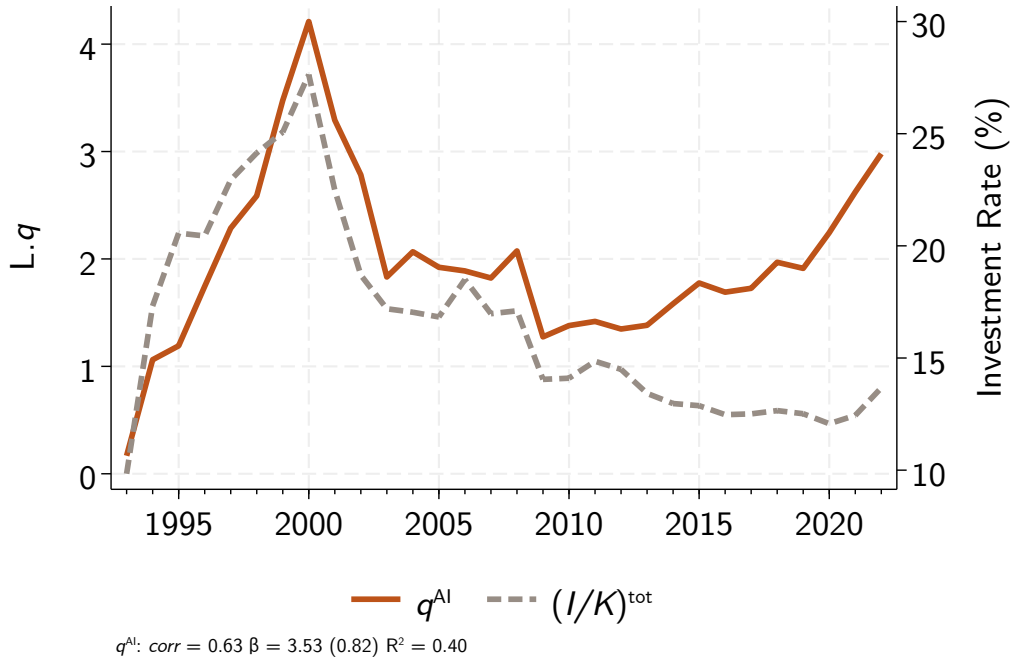
Tobin, J. 1969. A general equilibrium approach to monetary theory. *Journal of Money, Credit and Banking* 1:15–29.

Whited, T. M., and G. Wu. 2006. Financial constraints risk. *The Review of Financial Studies* 19:531–59.

Woeppel, M. 2022. Using patent capital to estimate tobin's q. *Journal of Financial and Quantitative Analysis* 57:2929–67.

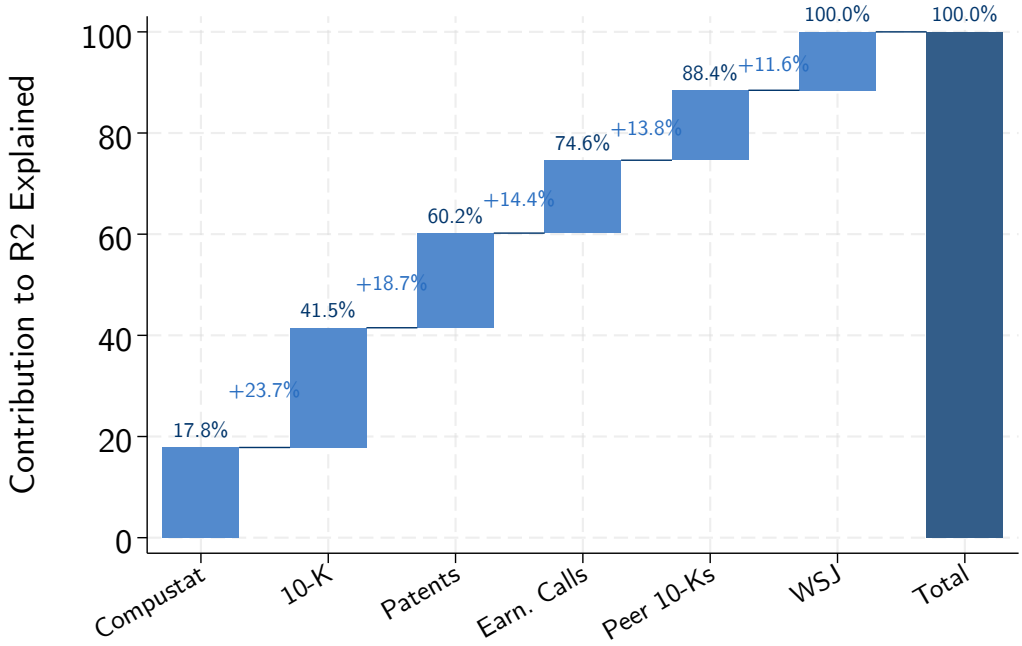
Figures

Figure 2. Aggregate Investment Rate and q Measures: Value-Weighted Time Series



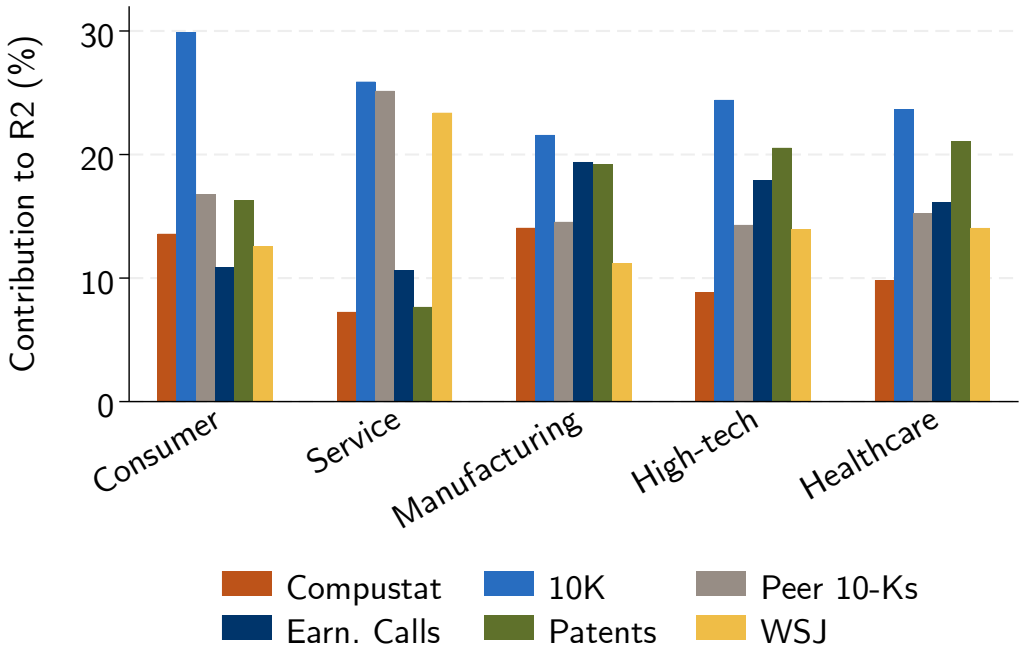
Notes. This figure plots the value-weighted aggregate investment rate (total investment scaled by total capital) alongside value-weighted aggregate q^{AI} over the sample period. Both series peak around the dot-com boom (~2000) and decline through the 2000s. Post-2010, q^{AI} recovers while the aggregate investment rate continues to drift downward, consistent with the "investmentless growth" pattern documented in Gutiérrez and Philippon (2017). The correlation between q^{AI} and the aggregate investment rate is 0.63 ($R^2 = 0.40$).

Figure 3. Shapley-Owen R^2 Decomposition of q^{AI}



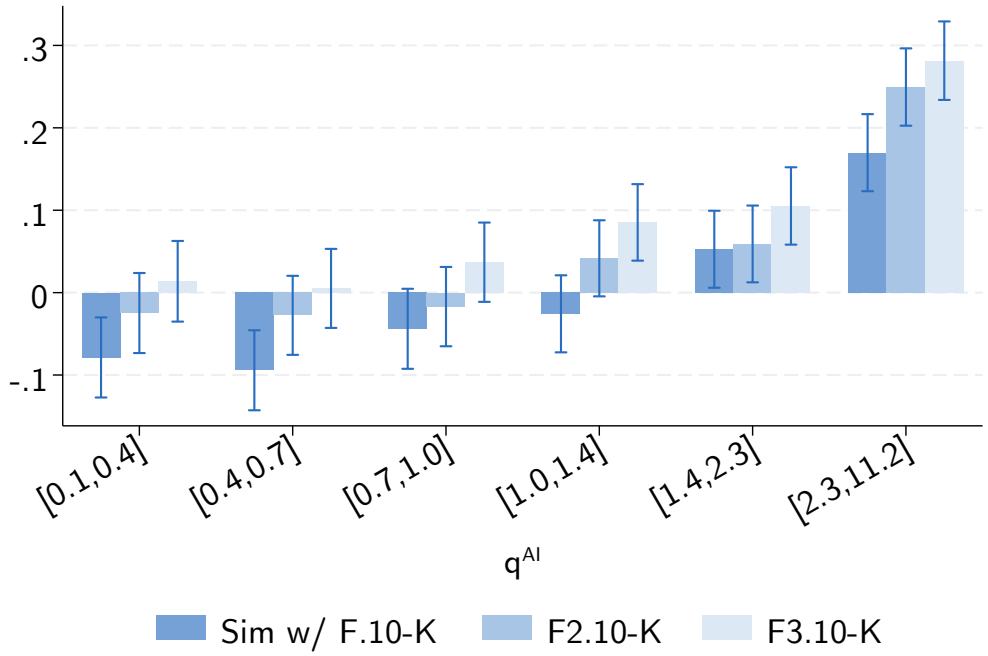
Notes. This figure reports a Shapley-Owen R^2 decomposition of q^{AI}, showing the marginal contribution of each information source to the variation in our AI-measured marginal q. Text sources collectively contribute more than 80% of q^{AI}'s variation. The firm's own 10-K filings are the single largest source (23.7%), followed by patents (18.7%), peer firm 10-Ks (13.8%), macroeconomic news from the Wall Street Journal (11.6%), and earnings conference calls (14.4%). Compustat financial variables alone explain only 17.8%. The decomposition demonstrates that q^{AI} draws primarily on textual evidence about the marginal project rather than repackaging accounting data.

Figure 4. Shapley-Owen R^2 Decomposition of q^{AI}: By Industry



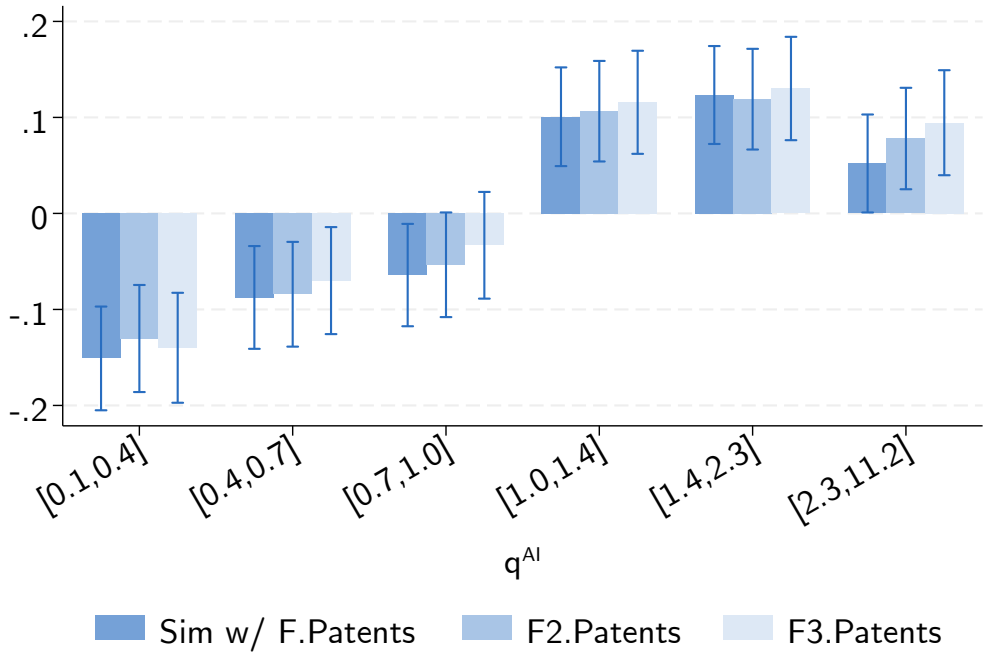
Notes. This figure reports the Shapley-Owen R^2 decomposition of q^{AI} separately by industry group. The composition of information sources varies meaningfully across industries. In high-tech and healthcare, patents contribute approximately 20% of the variation, reflecting innovation-driven investment opportunities. In manufacturing, contributions are most balanced across sources. In services, peer firm 10-Ks are unusually important (~25%) while patents are negligible, and Compustat accounting variables contribute the least (~7%). These patterns provide direct evidence that the measure draws on economically relevant textual content specific to each industry's investment opportunity set.

Figure 5. Validation: Similarity Between Generated Projects and Future 10-K Filings



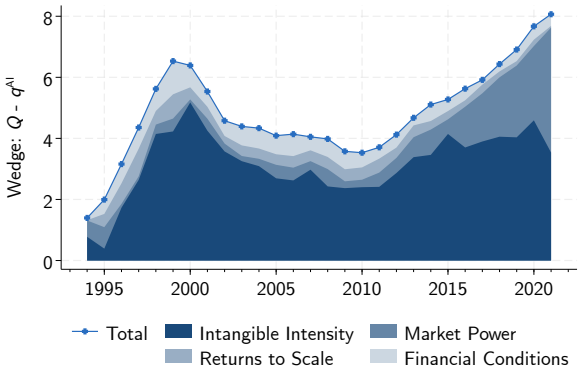
Notes. This figure plots the cosine similarity between LLM-generated project descriptions and the firm's *future* 10-K filings (Items 1 and 7) across multiple forward horizons, binned by q^{AI} quintiles. Higher q^{AI} firms generate projects that are significantly more similar to what the firm actually describes in subsequent annual reports. The monotonically increasing pattern across quintiles provides direct evidence that q^{AI} captures economically meaningful variation in forward-looking investment opportunities: firms with higher AI-measured marginal q go on to pursue projects more closely resembling those the system proposed.

Figure 6. Validation: Similarity Between Generated Projects and Future Patent Filings

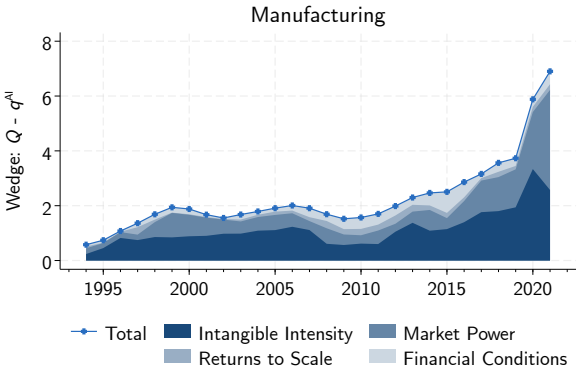


Notes. This figure plots the cosine similarity between LLM-generated project descriptions and the firm's *future* patent filings across multiple forward horizons, binned by q^{AI} quintiles. The pattern mirrors that of 10-K filings (Figure 5): firms with higher q^{AI} subsequently file patents that are more similar to the generated project descriptions. This validation along the innovation dimension confirms that q^{AI} captures information about the firm's R&D pipeline and technological direction, not just its broad strategic plans.

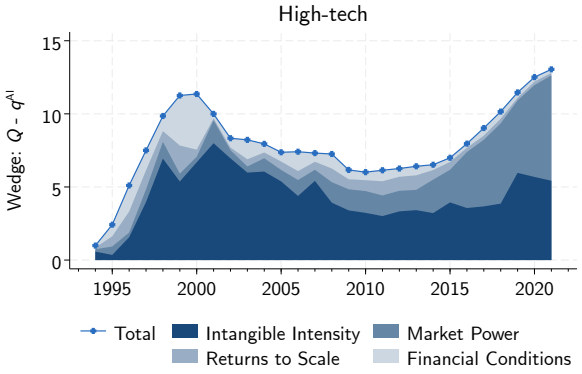
Figure 7. Decomposition of the Average Q–Marginal q^{AI} Wedge Over Time



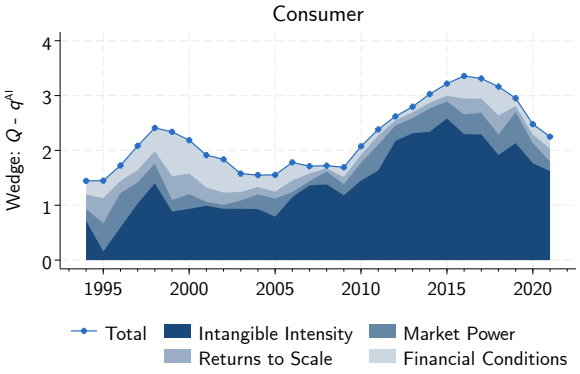
(a) All Industries



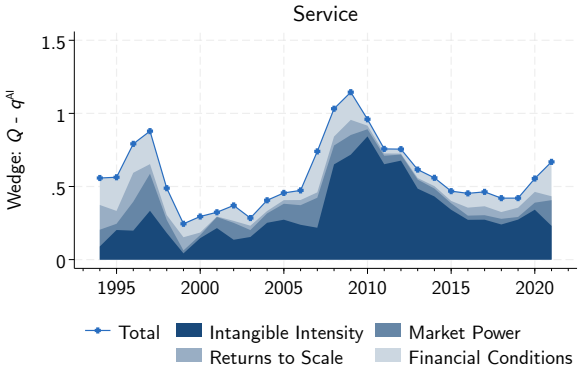
(b) Manufacturing



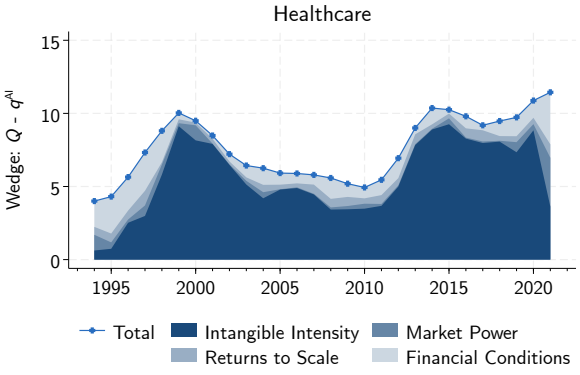
(c) High-Tech



(d) Consumer



(e) Services



(f) Healthcare

Notes. This figure decomposes the wedge between average Q and marginal q^{AI} into contributions from capitalized rents (profit margin), omitted intangibles (intangible intensity), returns to scale (total capital stock), and financial frictions (WW index, book leverage), plotted over time.

Tables

Table 2. Descriptive Statistics

	N	Mean	SD	25%	50%	75%
Physical Investment Rate (%)	10,616	4.400	4.516	1.704	3.065	5.289
Intangible Investment Rate (%)	10,616	12.226	9.222	6.150	10.020	15.543
R&D Investment Rate (%)	10,616	5.042	6.785	0.553	2.451	7.016
M&A Investment Rate (%)	10,616	3.427	8.943	0.000	0.000	1.903
Standard Investment Rate (%)	10,616	11.714	11.381	5.518	8.416	13.635
Total Investment Rate (%)	10,616	16.731	11.126	9.603	13.988	20.105
Q^{std}	10,616	8.977	47.713	1.259	2.977	7.027
Q^{tot}	10,616	1.645	2.480	0.580	1.043	1.816
q^{AI}	10,616	1.562	1.750	0.587	1.039	1.798
Total FCF (%)	10,616	13.069	14.186	6.663	12.173	19.020
Physical Capital	10,616	5,188	25,586	106	540	2,248
Intangible Capital	10,616	5,550	18,593	164	769	3,005
Total Capital	10,616	10,711	38,586	334	1,499	5,562
Market Value	10,616	14,587	63,806	272	1,433	7,227
1(Intangible)	10,616	0.714	0.276	0.667	0.667	1.000
1(R&D)	10,616	0.206	0.228	0.000	0.333	0.333
1(Buy)	10,616	0.125	0.184	0.000	0.000	0.333
Profit Margin (%)	10,616	41.142	20.653	27.018	39.367	54.226
Intangible Intensity (%)	10,616	59.469	22.894	42.266	62.245	78.238
WW Index	10,616	0.000	1.000	−0.662	−0.115	0.524
Book Leverage	10,616	0.232	0.216	0.055	0.205	0.338

Notes. This table reports the descriptive statistics for the firm-year panel used in the analysis.

Table 3. Investment-q Relation: Comparing Q^{std}, Q^{tot}, and q^{AI}

	(1)	(2)	(3)	(4)	(5)
	Investment Rate (investment scaled by total capital, %)				Standard Investment Rate (%)
	Physical	Intangible	Total	R&D	CAPX/PPE
L.Q^{std}	0.004	0.008	0.012	0.005	0.038**
	(0.003)	(0.007)	(0.009)	(0.005)	(0.017)
L.Q^{tot}	0.525***	0.855***	1.349***	0.546***	1.461***
	(0.058)	(0.120)	(0.184)	(0.087)	(0.222)
L.q^{AI}	0.911***	1.591***	2.552***	0.947***	2.648***
	(0.076)	(0.112)	(0.148)	(0.098)	(0.202)
ΔR^2 (tot − std)	0.137***	0.139***	0.168***	0.127***	0.105***
	(0.016)	(0.024)	(0.022)	(0.036)	(0.035)
ΔR^2 (AI − std)	0.147***	0.169***	0.209***	0.139***	0.128***
	(0.013)	(0.021)	(0.024)	(0.035)	(0.025)
ΔR^2 (AI − tot)	0.011	0.030**	0.042***	0.012	0.023*
	(0.011)	(0.013)	(0.015)	(0.008)	(0.012)
Firm FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Observations	10,616	10,616	10,616	10,616	10,616
No. of Firms	455	455	455	455	455
Mean Dep. Var.	4.399	12.227	16.731	5.043	11.712

Notes. This table reports panel regressions of firm-level investment rates on lagged q measures with firm and year fixed effects. Each panel row estimates a separate univariate regression of the investment rate on one of three q measures: standard average Q (Q^{std}), intangible-adjusted total Q (Q^{tot}, Peters and Taylor, 2017), and our AI-measured marginal q (q^{AI}). Columns (1)–(4) measure the investment rate as the ratio of investment to total capital (physical plus intangible); column (5) uses the standard investment rate (capital expenditures scaled by net PPE). The bottom panel reports pairwise differences in within-firm R^2 across the three q measures, with standard errors in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 4. Investment-q Relation: Controlling for Cash Flow

	(1)	(2)	(3)	(4)	(5)
	Investment Rate (investment scaled by total capital, %)				Standard Investment Rate (%)
	Physical	Intangible	Total	R&D	CAPX/PPE
Panel A: Without cash flow control					
L.q^{AI}	0.911***	1.591***	2.552***	0.947***	2.648***
	(0.076)	(0.112)	(0.148)	(0.098)	(0.202)
R^2	0.536	0.725	0.649	0.756	0.406
Panel B: With cash flow control					
L.q^{AI}	0.720***	1.350***	2.119***	0.952***	2.215***
	(0.076)	(0.137)	(0.176)	(0.107)	(0.222)
Total Cash Flow	0.070***	0.088***	0.158***	−0.002	0.159***
	(0.008)	(0.020)	(0.025)	(0.014)	(0.025)
R^2	0.555	0.732	0.665	0.756	0.421
Firm FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Observations	10,616	10,616	10,616	10,616	10,616
No. of Firms	458	458	458	458	458
Mean Dep. Var.	4.399	12.227	16.731	5.043	11.712

Notes. This table reports panel regressions of firm-level investment rates on lagged q^{AI} with and without a cash flow control, including firm and year fixed effects. Panel A replicates the q^{AI} specification from Table 3. Panel B adds total cash flow (scaled by total capital) as a regressor. Columns (1)–(4) measure the investment rate as the ratio of investment to total capital; column (5) uses capital expenditures scaled by net PPE. The cash flow coefficient is positive and significant for physical, intangible, and total investment, but shrinks substantially relative to specifications using average Q, consistent with the prediction of Abel (2018) that better measurement of q drives out the spurious explanatory power of cash flow. Cash flow is insignificant for R&D investment. Standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 5. Investment-q Relation: Measurement Error Correction

	(1)	(2)	(3)	(4)	(5)
	Investment Rate (investment scaled by total capital, %)				Standard Investment Rate (%)
	Physical	Intangible	Total	R&D	CAPX/PPE
Panel A: Q^{tot}					
L.Q^{tot}	1.142***	2.296***	2.955***	1.790***	3.153***
	(0.187)	(0.236)	(0.271)	(0.225)	(0.365)
ρ^2	0.195	0.268	0.318	0.262	0.292
	(0.058)	(0.049)	(0.052)	(0.059)	(0.062)
τ^2	0.442	0.626	0.652	0.545	0.551
	(0.098)	(0.104)	(0.120)	(0.103)	(0.108)
Panel B: q^{AI}					
L.q^{AI}	2.403***	3.814***	4.854***	3.028***	6.699***
	(0.139)	(0.169)	(0.189)	(0.180)	(0.247)
ρ^2	0.271	0.332	0.387	0.327	0.441
	(0.037)	(0.030)	(0.030)	(0.038)	(0.037)
τ^2	0.301	0.611	0.639	0.516	0.400
	(0.038)	(0.053)	(0.052)	(0.053)	(0.034)
Panel C: Difference (AI − tot)					
$\Delta\rho^2$ (AI − tot)	0.076	0.064**	0.069**	0.065*	0.148***
	(0.065)	(0.031)	(0.033)	(0.034)	(0.051)
$\Delta\tau^2$ (AI − tot)	−0.141	−0.015	−0.014	−0.028	−0.151**
	(0.325)	(0.075)	(0.051)	(0.054)	(0.064)
Firm FE	Y	Y	Y	Y	Y
Observations	10,616	10,616	10,616	10,616	10,616
No. of Firms	458	458	458	458	458
Mean Dep. Var.	4.399	12.227	16.731	5.043	11.712

Notes. This table reports higher-order cumulant estimators of the investment-q relation following Erickson, Jiang, and Whited (2014). Panel A reports results using intangible-adjusted total Q (Q^{tot}); Panel B uses q^{AI}. For each q measure, the table reports the bias-corrected coefficient, ρ^2 (the signal-to-total variance ratio, measuring how much of the variation in the q proxy reflects true investment opportunities), and τ^2 (the reliability ratio). Panel C reports the pairwise difference in ρ^2 and τ^2 between q^{AI} and Q^{tot}. The measurement-error-corrected coefficients for q^{AI} are substantially larger than the OLS estimates in Table 3, and q^{AI} exhibits higher ρ^2 than Q^{tot} across all investment types, confirming that our measure retains less attenuation bias than conventional proxies. Standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 6. Investment-q Relation: Heterogeneity by Project Type

	(1)	(2)	(3)
	Investment Rate (investment scaled by total capital, %)		
	Physical	Intangible	R&D
L.q^{AI}	0.854***	1.435***	0.883***
	(0.025)	(0.059)	(0.026)
L.**1**(Project Type)	−0.110	−0.069	−0.261**
	(0.100)	(0.154)	(0.119)
L.q^{AI} × L.**1**(Project Type)	0.132***	0.145**	0.228***
	(0.043)	(0.064)	(0.050)
Firm FE	Y	Y	Y
Year FE	Y	Y	Y
Observations	10,616	10,616	10,616
No. of Firms	458	458	458
R^2	0.537	0.730	0.765
Mean Dep. Var.	4.399	12.227	5.043

Notes. This table tests whether the investment-q elasticity varies with the type of marginal project identified by the LLM. In each column, the dependent variable is a type-specific investment rate (physical, intangible, or R&D investment scaled by total capital). **1**(Project Type) is an indicator equal to one if the LLM-generated marginal project is classified as matching the column's investment type, based on keyword analysis of the project description. The interaction term $q^{AI} \times$ **1**(Project Type) is positive and significant across all three columns, indicating that when the LLM identifies a physical (intangible, R&D) marginal project, the corresponding type of investment is more responsive to q^{AI}. All specifications include firm and year fixed effects. Standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 7. Investment-q Relation: Extensive versus Intensive Margin

	(1)	(2)	(3)	(4)	(5)
	Investment Rate (investment scaled by total capital, %)				Standard Investment Rate (%)
	Physical	Intangible	Total	R&D	CAPX/PPE
L.q^{AI}	0.772***	1.113***	1.882***	0.510***	2.124***
	(0.115)	(0.143)	(0.177)	(0.103)	(0.248)
L.q^{AI} × L.**1**(High Compatibility)	0.124	0.488***	0.679***	0.471***	0.547**
	(0.109)	(0.136)	(0.173)	(0.095)	(0.268)
Firm FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Observations	10,616	10,616	10,616	10,616	10,616
No. of Firms	458	458	458	458	458
R^2	0.537	0.731	0.648	0.767	0.400
Mean Dep. Var.	4.399	12.227	16.731	5.043	11.712

Notes. This table tests whether the investment-q elasticity depends on the compatibility of the marginal project with the firm's existing operations, following the extensive-versus-intensive margin framework of Jovanovic and Rousseau (2014). **1**(High Compatibility) is an indicator equal to one if the cosine similarity between the LLM-generated project description and the firm's contemporaneous 10-K filings and conference call transcripts exceeds the sample median. The interaction term $q^{AI} \times$ **1**(High Compatibility) is positive and significant for intangible, total, R&D, and standard investment, indicating that the investment-q elasticity is substantially larger when the marginal project is compatible with existing operations. This is consistent with reallocation frictions compressing the marginal response when firms must reorient their operations toward extensive-margin projects. All specifications include firm and year fixed effects. Standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 8. Investment-q Relation: Buy versus Build

	(1)	(2)	(3)	(4)
		Investment Rate (investment scaled by total capital, %)		
	1(M&A) (%)	M&A	Total	Total
L.**1**(Buy)	4.672*			
	(2.733)			
L.q^{AI}		2.340***	3.475***	2.459***
		(0.346)	(0.245)	(0.193)
Firm FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Sample	Full	Any M&A	Any M&A	No M&A
Observations	10,616	4,876	4,876	5,740
No. of Firms	458	381	381	420
R^2	0.397	0.213	0.715	0.636
Mean Dep. Var.	46.309	7.446	15.587	17.691

Notes. This table distinguishes between "building" and "buying" on the margin, following the q-theories of mergers in Jovanovic and Rousseau (2002) and Jovanovic and Rousseau (2003). **1**(Buy) is an indicator equal to one if the LLM-generated marginal project is classified as acquisition-based, using keyword analysis of the project description. Column (1) shows that "buy" projects predict actual M&A activity: the probability of undertaking an acquisition increases by 4.7 percentage points when the LLM identifies a buy-type marginal project. Columns (2)–(3) restrict the sample to firm-years with at least one acquisition and show that q^{AI} predicts both M&A-specific and total investment among acquirers. Column (4) restricts to firm-years with no M&A and confirms that q^{AI} remains a strong predictor of total investment among non-acquirers. All specifications include firm and year fixed effects. Standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 9. Dissecting the Average Q–Marginal q^{AI} Wedge

	(1)	(2)	(3)	(4)	(5)
	Wedge: $Q - q^{AI}$				
Profit Margin (%)	0.064***				0.053***
	(0.013)				(0.012)
Intangible Intensity (%)		0.126***			0.127***
		(0.013)			(0.012)
Log(K^{tot})			−0.335		−1.310***
			(0.259)		(0.265)
WW Index				−1.543***	−2.282***
				(0.325)	(0.327)
Book Leverage				−0.666	−0.018
				(0.850)	(0.781)
Firm FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
R^2	0.681	0.692	0.672	0.680	0.715

Notes. This table decomposes the wedge between average Q and marginal q^{AI} into three economically distinct sources. The dependent variable is $Q - q^{AI}$. Profit margin proxies for capitalized market-power rents (Crouzet and Eberly, 2023): higher profitability inflates average Q beyond what investment opportunities alone would justify. Intangible intensity (intangible investment share of total investment) captures omitted intangible capital: firms with greater intangible intensity have larger unmeasured capital stocks, widening the gap between average Q and the true marginal return. Log total capital stock and financial constraint measures (WW index, book leverage) capture scale-related frictions and decreasing returns to scale. Columns (1)–(4) enter each channel separately; column (5) includes all regressors jointly. Profit margin and intangible intensity are the dominant drivers of the wedge, with financial constraints (WW index) also significant. All specifications include firm and year fixed effects. Standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Internet Appendix

Putting Marginal Back in Tobin's q

Table of Contents

A Appendix Section **A2**

B Appendix Figures **A3**

C Appendix Tables **A4**

D Data Infrastructure and Text Processing Details **A8**

D.1 Raw Text Formats and Corpus-Specific Parsing A8

D.2 Unified Document Abstraction A9

D.3 Metadata Enrichment A9

D.4 Text Cleaning and Normalization A10

D.5 Chunking Procedure A10

D.6 Embedding Model A11

D.7 Vector Storage: Index Configuration and Namespaces A11

D.8 Batched Ingestion, Checkpointing, and Fault Tolerance A12

D.9 Ingestion Pipeline: Algorithmic Summary A12

E Methodology: Technical Details **A13**

E.1 Retriever Implementation A13

E.2 Retrieval Schedule and Per-Corpus Budgets A14

E.3 Q1 Retrieval Algorithm A15

E.4 Post-Retrieval Compression Details A16

E.5 Project Generation Details A17

E.6 Output Parsing Details A18

E.7 Supervised Fine-Tuning Details A19

F Example: Varian Medical Systems **A23**

F.1 Retrieved and Compressed Evidence A23

F.2 Generated Projects A25

F.3 q Computation and the Average–Marginal Wedge A27

F.4 Ex-Post Validation A28

A Appendix Section

B Appendix Figures

C Appendix Tables

Table A1. Keywords for Classifying LLM-Generated Project Types

Group	Keywords / patterns
Panel A: Physical Capital	
Plant & facility	plant, factor(y/ies), facilit(y/ies), warehouse, distribution center, fulfillment center, data center, infrastructure, construct(ion), build a new [plant/facility/warehouse/hub]
Machinery & equipment	machinery, equipment, apparatus, instrument(s), device(s), sensor(s)
Real estate & land	real estate, property acquisition/purchase/development, land acquisition/purchase
Logistics	fleet, vessel(s), aircraft, truck(s)
Capacity expansion	production capacity, manufacturing capacity, scale-up production/manufacturing/output, expand production/manufacturing capacity
Energy & utilities	power plant, renewable energy project/asset, solar farm/panel/plant, wind farm/turbine, pipeline, drilling, mine/mines/mining, extractive
Panel B: Intangible Capital	
Software, platform & digital	software, platform, app, application, digital, cybersecurity, cloud platform/service/infrastructure/computing, AI, artificial intelligence, machine learning, algorithm(s), analytics, data platform/analytics/science/pipeline/asset
IP & patents	patent(s/ed/ing), intellectual property, IP, know-how, trade secret(s), copyright(s), trademark(s)
R&D / scientific	research and development, R&D, preclinical, clinical trial(s), drug development/discovery/candidate, therapeutic(s), formula(tion)
Brand & marketing	brand, brand equity/awareness/recognition/value, marketing, advertising, customer relationship/acquisition/base/loyalty
Human capital	human capital, workforce, talent, training program(s), skills development
Organizational & process	business process, organizational, management system, supply-chain optimization/management/platform, licensing technology/IP/patent
Panel C: R&D	
Core R&D terms	research and development, R&D, basic research, applied research, scientific research, research program(s)
Experimentation & prototyping	experiment(al/ation), prototyp(e/ing), pilot project/program/study/test, feasibility study, proof-of-concept, PoC
Clinical & biomedical	preclinical, clinical trial(s), Phase I/II/III trial/study, drug development/discovery/candidate
Invention & innovation	invent(ion/or), innovat(e/ion), patent develop.../appli.../fil..., technology development, software development, platform development
Advanced & next-generation	new technology, next-gen(eration), advanced material/technology/compound/process, breakthrough, pioneer(ing)

Continued on next page

Table A1 (continued)

Group	Keywords / patterns
Panel D: Buy versus Build	
Buy keywords:	
M&A & takeover	acquir(e/es/ing), acquisition(s), merger(s), merging, takeover, take over, buyout, buy out, absorb(ing/ption), consolidat(e/ion)
Purchase	purchas(e/ing/ed), procur(e/ement)
Licensing-in	license/licens... from, in-licens(e/ing), licensing rights
Outsourcing	outsourc(e/ing)
Build keywords:	
R&D & development	develop(ing/ment/ed), research(ing), R&D, innovat(e/ion/ive), invent(ion/or)
Physical construction & manufacturing	build(ing/s), construct(ion), manufactur(e/ing), new facilit(y/ies), new plant, new production line, establish... facilit(y/ies), scale up, expand... capacity, expand... facilit(y/ies), expand... production capacity
Organic launch & commercialization	launch(ing/ed), introduc(e/ing/tion), commercializ(e/ation), pilot(ing), prototyp(e/ing), implementing
IP creation	patent(s/ing/ed), design(ing)
Licensing-out	licens... [technology] to, offering licensing

Notes. This table lists the keyword dictionaries used to classify LLM-generated project descriptions into investment types. For each firm-year, we apply these keywords to the text of the three generated project descriptions. Panel A classifies projects as involving physical capital investment. Panel B classifies projects as involving intangible capital. Panel C identifies R&D-intensive projects. Panel D distinguishes between "buy" projects (acquisitions, purchases, licensing-in) and "build" projects (organic development, construction, commercialization). Parenthetical notation indicates morphological variants matched by the classifier (e.g., acquir(e/es/ing) matches acquire, acquires, and acquiring). A project may be assigned to multiple categories if its description contains keywords from more than one panel.

Table A2. Investment-q Relation: Horse Race

	(1)	(2)	(3)	(4)	(5)
	Investment Rate (investment scaled by total capital, %)				Standard Investment Rate (%)
	Physical	Intangible	Total	R&D	CAPX/PPE
Panel A: q^{AI} alone					
L.q^{AI}	0.911***	1.591***	2.552***	0.947***	2.648***
	(0.076)	(0.112)	(0.148)	(0.098)	(0.202)
R^2	0.536	0.725	0.649	0.756	0.406
Panel B: Q^{tot} alone					
L.Q^{tot}	0.525***	0.855***	1.349***	0.546***	1.461***
	(0.058)	(0.120)	(0.184)	(0.087)	(0.222)
R^2	0.528	0.713	0.625	0.752	0.387
Panel C: q^{AI} and Q^{tot} jointly					
L.q^{AI}	0.618***	1.279***	2.121***	0.640***	2.005***
	(0.089)	(0.181)	(0.225)	(0.143)	(0.310)
L.Q^{tot}	0.231***	0.246**	0.340***	0.242**	0.506**
	(0.042)	(0.111)	(0.127)	(0.105)	(0.241)
R^2	0.541	0.726	0.650	0.758	0.409
Firm FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Observations	10,616	10,616	10,616	10,616	10,616
No. of Firms	455	455	455	455	455
Mean Dep. Var.	4.399	12.227	16.731	5.043	11.712

Notes. This table reports horse-race regressions in which q^{AI} and Q^{tot} are entered jointly. Panel A reproduces the univariate q^{AI} specification from Table 3; Panel B reproduces the univariate Q^{tot} specification. Panel C enters both measures simultaneously. When both are included, the coefficient on q^{AI} declines modestly (e.g., from 2.552 to 2.121 for total investment) but remains large and highly significant, while the coefficient on Q^{tot} drops substantially (from 1.349 to 0.340) and retains significance at a reduced level. The R^2 in the joint specification is only marginally above the q^{AI}-only R^2, indicating that Q^{tot} contributes little incremental explanatory power once q^{AI} is controlled for. All specifications include firm and year fixed effects. Standard errors clustered by firm are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A3. Investment-q Relation: Obviously Related Instrumental Variables

	(1)	(2)	(3)	(4)	(5)
	Investment Rate (investment scaled by total capital, %)				Standard Investment Rate (%)
	Physical	Intangible	Total	R&D	CAPX/PPE
Panel A: q^{AI} (OLS)					
L.q^{AI}	0.911***	1.591***	2.552***	0.947***	2.648***
	(0.076)	(0.112)	(0.148)	(0.098)	(0.202)
Panel B: Q^{tot} (OLS)					
L.Q^{tot}	0.525***	0.855***	1.349***	0.546***	1.461***
	(0.058)	(0.120)	(0.184)	(0.087)	(0.222)
Panel C: ORIV					
L.Corrected q	0.934***	1.564***	2.486***	0.971***	2.645***
	(0.079)	(0.119)	(0.147)	(0.116)	(0.243)
Firm FE	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Observations	10,616	10,616	10,616	10,616	10,616
No. of Firms	455	455	455	455	455
Mean Dep. Var.	4.399	12.227	16.731	5.043	11.712

Notes. This table reports Obviously Related Instrumental Variables (ORIV) estimates following the approach of Gillen, Snowberg, and Yariv (2019). The ORIV estimator treats q^{AI} and Q^{tot} as two noisy proxies for the same latent variable (true marginal q) and uses each as an instrument for the other in a stacked IV regression. Panels A and B reproduce the OLS benchmarks from Table 3. Panel C reports the ORIV estimates. The ORIV coefficients are close to the q^{AI} OLS estimates (e.g., 2.486 vs. 2.552 for total investment) and substantially above the Q^{tot} OLS estimates, consistent with q^{AI} being the less-attenuated proxy. All specifications include firm, year, and group (stacking) fixed effects. Standard errors clustered by firm are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

D Data Infrastructure and Text Processing Details

This appendix provides technical details on the text-to-vector pipeline underlying the data infrastructure described in Section 2.3. These details are intended for readers interested in replicating or extending the text processing architecture. The main text describes the economic motivation for each design choice; here we document the implementation.

D.1 Raw Text Formats and Corpus-Specific Parsing

Each text corpus arrives in a different raw format and requires source-specific parsing before it can be processed by the unified embedding pipeline.

SEC 10-K filings. We obtain 10-K filings from SEC EDGAR as parsed JSON files. Each JSON file corresponds to a single filing, with keys representing section titles (e.g., `Item 1`, `Item 7`) and values containing the corresponding section texts. Filings are organized in a directory tree and discovered via recursive scanning. We extract filing metadata—including CIK, filing year, form type, and accession number—deterministically from the filename, which follows the convention:

`date_formType_edgar_data_cik_accession.json`.

We create one document per section, storing the section title as a metadata field. For our primary analysis, we retain only `Item 1` and `Item 7`, separating them either by metadata filter or by assigning them to distinct vector store namespaces.

Patent filings. Each patent record is treated as a text document associated with firm-year identifiers. We store patents with `gvkey` and `filingyear` as primary keys, along with descriptive fields including patent title and, when available, a patent identifier. The `gvkey` field is cast to string format to ensure compatibility across ingestion pipelines.

Earnings conference calls. Raw transcripts are stored in Parquet format with fields including the full transcript text, `CUSIP`, `Fiscal Year`, and `Latest Call Date`. We drop records with

missing identifiers or empty text. Each retained transcript is converted into a single document for downstream processing.

WSJ front pages. Each front-page article is stored as a record containing, at minimum, a `title`, a `year`, and the article `text`. We treat each article as a single document and store embeddings in a dedicated WSJ namespace for macro-context retrieval.

D.2 Unified Document Abstraction

Despite the heterogeneity of raw formats, all corpora are standardized into a common document representation before entering the embedding pipeline. Each document consists of two components:

1. `page_content`: the raw text string to be embedded;
2. `metadata`: a dictionary of identifiers and descriptors that enables downstream filtering and traceability.

This abstraction allows a single chunking, embedding, and indexing pipeline to operate across all corpora with minimal source-specific code.

D.3 Metadata Enrichment

We enrich raw metadata to support firm-year and industry-level filtering in downstream retrieval. The metadata schema for each corpus is designed so that all documents, regardless of source, can be linked to a common firm-year key.

CIK-to-firm mapping for 10-K filings. SEC filings are natively identified by CIK (Central Index Key), which does not directly correspond to the Compustat firm identifiers (`gvkey`) used in our financial panel. We merge an externally prepared crosswalk from CIK to `gvkey`, SIC industry code (`sich`), and company name (`conm`). When a company name is unavailable in the crosswalk, we assign a deterministic placeholder derived from the CIK to maintain completeness. This enrichment is essential for two retrieval operations: (i) restricting 10-K retrieval to the focal firm's own filings (filtering on `gvkey`), and (ii) retrieving peer firm disclosures from the same industry (filtering on `sich`).

Cross-corpus identifier alignment. Our downstream unit of observation is the firm-year. Accordingly, the metadata schema for each corpus is designed to support linking to a common firm-year key:

- 10-K filings: `cik`, and (via the CIK crosswalk) `gvkey` and `fyear`;
- Patents: `gvkey` and `filingyear`, plus `patent_title` and optionally `patent_id`;
- Conference calls: `CUSIP` and `fiscal_year`, plus `latest_call_date`;
- WSJ: `year` and `title`.

This design allows retrieval to be conditioned jointly on firm identity, time period, and source type through metadata-filtered vector search.

D.4 Text Cleaning and Normalization

Before chunking and embedding, we apply a uniform text cleaning procedure to reduce artifacts from upstream extraction and to stabilize segment boundaries. The cleaning pipeline performs three operations: (i) removal of page-number patterns commonly produced by PDF-to-text conversion (e.g., isolated digits following line breaks); (ii) replacement of newline characters with single spaces; and (iii) collapsing of repeated whitespace into single spaces. We apply the same normalization logic across all corpora (10-K, patents, conference calls, WSJ) to ensure that the embedding model processes consistently formatted text.

D.5 Chunking Procedure

Long documents must be split into smaller units to be embedded and retrieved effectively. We use a recursive character-based splitting procedure to partition each cleaned document into overlapping chunks. Specifically, given a cleaned text T, we generate a sequence of chunks

$$C_1, C_2, \ldots, C_K, \quad \text{where } |C_k| \leq 512 \text{ characters and } C_k \cap C_{k+1} \approx 20 \text{ characters.}$$

The 20-character overlap mitigates boundary effects: semantically related sentences that straddle a split point appear in at least one chunk. Each chunk inherits the full metadata of its parent

document, preserving traceability from vector record to source document to firm-time identifiers.

Caching. To avoid recomputing the deterministic chunking step during iterative development, we serialize the list of chunked documents to disk and reuse the cached result when the embedding pipeline is re-run. This improves reproducibility and reduces compute cost.

D.6 Embedding Model

We embed each text chunk using `multilingual-e5-large` (`intfloat/multilingual-e5-large` on Hugging Face), a pre-trained sentence embedding model based on a 24-layer transformer architecture. Embeddings are computed synchronously during batch ingestion. For each chunk C_k, the model returns a fixed-length vector $\mathbf{v}_k \in \mathbb{R}^{1024}$. The model is trained on approximately one billion text pairs spanning 100 languages, with the training objective ensuring that semantically similar texts are mapped to nearby vectors (as measured by cosine similarity). We choose this model for its strong performance on semantic similarity and retrieval benchmarks, as well as its robustness to the heterogeneous language styles present across our corpora—ranging from legal prose in 10-K filings to natural speech in conference calls.

D.7 Vector Storage: Index Configuration and Namespaces

We store all embeddings in a Pinecone serverless vector index configured with cosine similarity as the distance metric, with the index dimension set to $d = 1{,}024$ to match the embedding model output. We organize the index into corpus-specific namespaces:

$$\{\texttt{10k_item1}, \texttt{10k_item7}, \texttt{patents}, \texttt{conference_call}, \texttt{wsj_frontpage}\}.$$

Namespace separation enables retrieval to be restricted by document type without text-based heuristics. Operationally, we run the same ingestion script once per corpus, changing only the data loader and the target namespace.

Each stored vector record consists of three components: (i) the embedding vector $\mathbf{v}_k$; (ii) the original chunk text C_k (stored as the document content field); and (iii) the full metadata dictionary inherited from the parent document. This design supports metadata-filtered semantic search—for

example, retrieving only 10-K Item 1 chunks from the focal firm's SIC industry, filed before a given year—while preserving the ability to inspect the underlying text of any retrieved result.

D.8 Batched Ingestion, Checkpointing, and Fault Tolerance

Given the scale of the chunked corpus (millions of text segments across all sources), we ingest embeddings into the vector index in batches of 1,000 chunks. To support fault tolerance in long-running ingestion jobs, we implement a checkpoint mechanism that records the index of the last successfully processed batch to a local file. When the pipeline is re-run—whether due to interruption or incremental data addition—it resumes from the checkpoint rather than restarting from the beginning. To handle transient failures (e.g., network timeouts), we wrap each batch insertion in a retry loop that waits briefly and retries until the batch completes before advancing the checkpoint.

D.9 Ingestion Pipeline: Algorithmic Summary

Algorithm 1 summarizes the end-to-end text-to-vector ingestion pipeline. The same algorithm is executed once per corpus, parameterized by the data loader and target namespace.

Algorithm 1 Unified text-to-vector ingestion pipeline

Require: Raw corpus files $\mathcal{F}$ for source s; namespace n_s; embedding model $E(\cdot)$; vector index I
Ensure: Index I contains chunk vectors for source s under namespace n_s

Load auxiliary mappings for metadata enrichment (e.g., CIK → SICH/CONM for 10-K)
Initialize vector store manager bound to (I, n_s)
Parse raw files into documents $\{(t_j, m_j)\}_{j=1}^{J}$, where t_j is text and m_j is metadata
Clean text: $t_j \leftarrow \textsc{Clean}(t_j)$; discard documents with empty text
Split into overlapping chunks: $\{(C_k, m_k)\}_{k=1}^{K} \leftarrow \textsc{Chunk}(\{(t_j, m_j)\})$
(Optional) Cache $\{(C_k, m_k)\}$ to disk for reuse
Load checkpoint: set starting batch index b_0
for each batch B of 1,000 chunks, starting from b_0 **do**
    Compute embeddings and insert: $\{(\mathbf{v}_k = E(C_k),\ C_k,\ m_k)\}_{k \in B} \rightarrow (I, n_s)$
    if insertion fails **then**
        Wait briefly and retry the same batch
    else
        Save checkpoint with completed batch index
    end if
end for

E Methodology: Technical Details

This appendix provides the complete technical specification of each stage of the LLM-RAG pipeline described in Section 3. The main text describes the economic motivation and conceptual design; here we document the implementation parameters, algorithm pseudocode, and engineering details needed for replication.

E.1 Retriever Implementation

For each corpus-specific retrieval call, we execute a five-step retriever pipeline that converts a natural-language query into a compact, non-redundant set of text passages. The retriever operates on a Pinecone vector index (described in Appendix D.7).

Step 1: Query embedding. Given a query string q, we compute its embedding $\mathbf{e}(q)$ using the same model as for indexing (`multilingual-e5-large`; see Appendix D.6), ensuring that query and document embeddings inhabit the same vector space.

Step 2: Widened vector search. We query Pinecone using cosine similarity, restricted to the relevant namespace and metadata filter. Rather than fetching exactly k candidates, we fetch a larger set to allow subsequent filtering and deduplication:

$$k_{\text{raw}} = k \times m_{\text{raw}},$$

where $m_{\text{raw}} = 10$ is the raw fetch multiplier. This widened window ensures that high-quality candidates are not lost to near-duplicate removal or minimum-length filtering.

Step 3: Minimum-length filter. We discard candidates whose stored chunk text is shorter than 300 characters, as these typically correspond to boilerplate headers, page breaks, or truncated sentences that provide little substantive evidence.

Step 4: Fuzzy near-duplicate removal. To reduce prompt bloat from overlapping chunks and repeated boilerplate, we remove near-duplicate passages using a fuzzy matching heuristic. Each

candidate text is canonicalized by lowercasing, Unicode normalization, ASCII folding, whitespace compression, and masking of four-digit years (which prevents superficial year differences from blocking otherwise-identical passages). We then compute RapidFuzz token-sort similarity against previously accepted candidates and reject chunks exceeding a similarity threshold of 90.

Step 5: Optional cross-encoder reranking. After filtering and deduplication, we retain up to $k \times m_{\text{cand}}$ candidates (with candidate multiplier $m_{\text{cand}} = 3$) and optionally rerank them using a cross-encoder model (default: `cohere-rerank-3.5`). Cross-encoder reranking jointly scores the query-document pair rather than comparing pre-computed embeddings, yielding more precise relevance judgments at the cost of additional inference. We apply reranking to at most 200 candidates for efficiency. If reranking fails (e.g., due to a timeout), we fall back to the original cosine similarity ordering.

Output formatting. The top-k passages are returned as a numbered list for readability and to facilitate citation in downstream generation.

E.2 Retrieval Schedule and Per-Corpus Budgets

Q1 retrieval is organized into three sequential stages with progressive query expansion, as described in Section 3.1. Table A4 reports the default per-corpus top-k budgets.

Table A4. Default per-corpus retrieval budgets for Q1.

Notes: Each budget specifies the number of text passages retained after deduplication and reranking. The retriever fetches $k \times 10$ raw candidates from the vector index before applying the filtering pipeline described in Appendix E.1.

Corpus (Namespace)	Top-k	Stage
10-K Item 1, self (`10k_item1`)	20	A
10-K Item 7, self (`10k_item7`)	20	A
10-K Item 1, peers (`10k_item1`)	20	B
10-K Item 7, peers (`10k_item7`)	20	B
Conference calls (`conference_call`)	90	C
Patents (`patents`)	10	C
WSJ front pages (`wsj_frontpage`)	45	C

Metadata filters. For each firm-year ($\texttt{gvkey}_0$, $\texttt{fyear}_0$), the retriever applies the following metadata filters to enforce temporal integrity and firm-level targeting:

$$\mathcal{F}_{\text{10K}}^{\text{self}} : \{\texttt{gvkey} = \texttt{gvkey}_0,\ \texttt{fyear} < \texttt{fyear}_0\}, \tag{A1}$$

$$\mathcal{F}_{\text{10K}}^{\text{peer}} : \{\texttt{sich} = \texttt{sich}_0,\ \texttt{gvkey} \neq \texttt{gvkey}_0,\ \texttt{fyear} < \texttt{fyear}_0\}. \tag{A2}$$

Analogous filters for other corpora use `filing_year` (patents), `fiscal_year` (conference calls), or `year` (WSJ), each with a strict less-than constraint on the focal fiscal year.

E.3 Q1 Retrieval Algorithm

Algorithm 2 summarizes the multi-stage Q1 retrieval and prompt construction procedure.

Algorithm 2 Q1 multi-source retrieval and prompt construction

Require: Firm-year ($\texttt{gvkey}_0$, $\texttt{fyear}_0$); master table with `conm`, `sich`, `cusip`; Pinecone index
Ensure: Q1 prompt containing multi-source retrieved evidence (prior to $\texttt{fyear}_0$)

Read identifiers ($\texttt{conm}_0$, $\texttt{sich}_0$, $\texttt{cusip}_0$) from the master table
Define metadata filters $\mathcal{F}_{\text{10K}}^{\text{self}}$ and $\mathcal{F}_{\text{10K}}^{\text{peer}}$ with strict $\texttt{fyear} < \texttt{fyear}_0$
Stage A: Retrieve focal firm disclosures
  `txt_item1` $\leftarrow$ RETRIEVE(base query, `10k_item1`, k_{item1}, $\mathcal{F}_{\text{10K}}^{\text{self}}$)
  `txt_item7` $\leftarrow$ RETRIEVE(base query$\|$`txt_item1`, `10k_item7`, k_{item7}, $\mathcal{F}_{\text{10K}}^{\text{self}}$)
  `txt_items` $\leftarrow$ `txt_item1`$\|$`txt_item7`
Stage B: Retrieve peer firm disclosures
  `txt_peers` $\leftarrow$ RETRIEVE(base query$\|$`txt_items`, peer namespaces, k_{peer}, $\mathcal{F}_{\text{10K}}^{\text{peer}}$)
Stage C: Retrieve auxiliary corpora
  $q_{\text{aux}} \leftarrow$ base query$\|$`txt_items`$\|$`txt_peers`
  `txt_pat` $\leftarrow$ RETRIEVE(q_{aux}, `patents`, k_{pat}, $\{\texttt{gvkey} = \texttt{gvkey}_0, \texttt{filing_year} < \texttt{fyear}_0\}$)
if $\texttt{cusip}_0$ is available **then**
  `txt_conf` $\leftarrow$ RETRIEVE(q_{aux}, `conference_call`, k_{conf}, $\{\texttt{CUSIP} = \texttt{cusip}_0, \texttt{fiscal_year} < \texttt{fyear}_0\}$)
end if
  `txt_wsj` $\leftarrow$ RETRIEVE(q_{aux}, `wsj_frontpage`, k_{wsj}, $\{\texttt{year} < \texttt{fyear}_0\}$)
Assemble the Q1 prompt by injecting all retrieved blocks into a fixed template

The primitive RETRIEVE$(\cdot)$ denotes the five-step retriever described in Appendix E.1.

E.4 Post-Retrieval Compression Details

Section-marker scaffold. The raw Q1 prompt is organized using machine-parsable section markers that delimit each evidence block. Each block is bounded by a `START` marker (e.g., `=== QUARTER/ANNUAL REPORTS FOR [company] ===`) and a corresponding `END` marker (e.g., `=== END OF REPORTS ===`). A dedicated `QUESTION` block preserves the task definition used to condition the section-wise summarization. This scaffold enables fully automated compression: a regular expression parser locates each block, extracts its text, and routes it to the summarizer independently.

Token budgeting. Before sending each evidence block to the summarizer, we compute a token budget to prevent context overflow. Let L_{max} denote the summarizer model's context limit and L_{out} the desired maximum output tokens. The available budget for the section text is:

$$L_{\text{sec}} = L_{\text{max}} - L_{\text{out}} - L_{\text{overhead}} - L_{\text{margin}},$$

where L_{overhead} is the token cost of system and user wrappers (estimated using the model's chat template) and L_{margin} is a safety buffer. If the raw section exceeds L_{sec}, we apply deterministic truncation (head truncation with a boundary marker) to ensure the request does not fail.

Summarization output schema. The summarizer is instructed to produce bullet-point output grouped under the following fixed headings: (i) Core business and strategy; (ii) Performance / MD&A signals; (iii) Current projects / investment focus / business model; (iv) Market / competition; (v) Risks / outlook; (vi) Other useful facts (optional). This schema is aligned with the analytical goals of the project generation stage.

Prompt revision. After generating summaries for all non-empty sections, the pipeline reconstructs the prompt by replacing each raw evidence block with its compressed summary, preserving the section markers and overall prompt structure. The result is a revised prompt with the same architecture as the original but substantially smaller and less redundant.

Large-scale execution. The compression is applied over the full firm-year panel. To support fault tolerance on cluster environments, we implement a resume mechanism (skip observations

Algorithm 3 Section-wise post-retrieval compression

Require: Raw Q1 prompt P_1^{raw} with section markers; tokenizer; LLM summarizer; $(L_{\max}, L_{\text{out}})$
Ensure: Revised prompt P_1^{rev} with compressed evidence blocks

```
Extract question text Q from the QUESTION block
for each section type t ∈ {company, peers, patents, calls, wsj} do
    Locate span (START_t, END_t) and extract section text S_t
    if S_t is empty then
        Set summary Ŝ_t ← ∅
    else
        Compute token budget L_sec and truncate S_t if needed
        Query LLM: Ŝ_t = SUMMARIZE(Q, S_t)
    end if
end for
Replace each S_t between markers with Ŝ_t to form P_1^rev
```

with existing non-empty output), optional sharding across parallel jobs (process rows where `row_idx` mod `num_shards` = `shard_id`), and automatic skipping of observations with malformed section markers.

E.5 Project Generation Details

Q2 prompt construction. The generation prompt P_2^{rev} is constructed from the compressed Q1 prompt P_1^{rev} by: (i) removing the Q1 question block (everything from the `=== QUESTION ===` marker onward), retaining only the evidence sections; (ii) appending a structured financial snapshot for the focal fiscal year (extracted from a precomputed JSON table); and (iii) appending a revised instruction that requires the model to output exactly three project objects in a strict JSON array format. Each project object must contain: `PROJECT`, `DESCRIPTION`, `MARKET VALUE` (USD million), `IMPLEMENTATION COST` (USD million), `REASONING`, `CONFIDENCE` (0–100), and `SIMILAR FIRMS` (three public comparables with tickers). The model is instructed to return only the JSON array, with no surrounding text or markdown formatting.

Serving protocol. We generate Q2 outputs using a vLLM server exposing an OpenAI-compatible endpoint (`/v1/chat/completions`). The generation script does not start the model server; instead, a batch scheduler (e.g., Slurm) launches the server and then invokes the client. Before processing, the client verifies server availability by querying `/v1/models`.

Context-length management. We use conservative decoding parameters (temperature close to zero) and cap generation length. For stability, we implement a guardrail: estimate prompt length in tokens using a bytes-to-token heuristic, shrink `max_tokens` if necessary to fit within the server's maximum model length minus a safety margin, and retry with halved `max_tokens` if the server returns a context-length error.

Incremental writing and resume. Because prompts and outputs are large multiline strings, we write the output CSV incrementally row by row, keyed by (`gvkey`, `fyear`), with proper quoting and escaping. A lightweight progress file (`.progress.tsv`) logs each completed observation for cluster monitoring. Resumed runs skip completed keys automatically.

Algorithm 4 Q2 prompt construction and large-scale generation

Require: Revised evidence prompt P_1^{rev}; financial snapshot JSON $J(\texttt{fyear})$; vLLM server
Ensure: Output CSV with (`gvkey`, `fyear`, P_2^{rev}, `q2_answer`)
Construct $P_2^{\text{rev}} \leftarrow \textsc{BuildQ2}(P_1^{\text{rev}}, J(\texttt{fyear}))$
Verify server availability via `/v1/models`
for each firm-year (`gvkey`, `fyear`) not yet completed **do**
    Choose `max_tokens` using a conservative context-length budget
    Call `/v1/chat/completions` with user message P_2^{rev}
    if context-length error **then**
        Shrink `max_tokens` and retry (bounded)
    end if
    Append one row to output CSV and one line to progress TSV
end for

E.6 Output Parsing Details

JSON recovery. Given a raw response string R, we apply the following deterministic transformations: (i) strip surrounding whitespace and remove optional markdown code fences; (ii) if non-JSON text surrounds the array, extract the substring from the first `[` to the last `]` (or from the first { to the last }); (iii) remove trailing-comma patterns. We then attempt standard JSON parsing; if that fails, we fall back to Python literal evaluation. If parsing still fails, we return an empty list.

Key normalization. To handle minor schema drift (e.g., `MARKET_VALUE` vs. `Market Value`), we normalize all keys by lowercasing and removing whitespace and underscores before field extraction.

Output artifacts. The post-processing stage produces two files: (i) a raw results file containing the prompt and raw model output for each firm-year, and (ii) a rectangular file that expands the JSON array into structured columns, including (q_1, q_2, q_3) computed from the extracted market values and implementation costs.

Table A5. Key intermediate artifacts in the Q1/Q2 pipeline.

Artifact	Contents / Role
`results_q1_q2.csv`	Raw Q1 prompt with retrieved sections and Q2 prompt template, keyed by (`gvkey`, `fyear`).
`results_q1_prompt_revised.csv`	Post-retrieval compressed Q1 prompts (section-wise summaries).
`results_q2_revised_with_answers.csv`	Revised Q2 prompts and raw model outputs.
`results_q2_revised_with_answers_split.csv`	Rectangular dataset with extracted project fields and computed q_i.

E.7 Supervised Fine-Tuning Details

E.7.1 SFT Sample Construction

The SFT sample is constructed by merging the revised Q2 results file (containing prompts and raw model outputs) with a firm-year panel containing the benchmark q_{tot}, using (`gvkey`, `fyear`) as the join key. We retain only observations for which the revised prompt, the raw Q2 answer, and q_{tot} are all observed and the answer parses into exactly three valid project objects.

For each retained observation, we compute baseline raw project-level ratios $q_i^{\text{raw}} = \texttt{MARKET VALUE}_i/\texttt{IMPLEMENT}$ as diagnostics. We then construct the pseudo-label completion by recalibrating the numerical fields as described in Section 3.5.1. If total investment cannot be parsed from the prompt, we fall back to the original implementation cost when it is available and positive; otherwise we assign a small positive fallback value to maintain a well-defined ratio.

E.7.2 Grouped Train/Validation/Test Splits

We partition the sample using grouped random splits to reduce leakage across related observations. The grouping variable is the firm identifier `gvkey` (alternatively, `gvkey_fyear` for robustness).

The default design assigns 80% of groups to training and 20% to a temporary holdout, then splits the holdout evenly into validation and test, yielding an 80/10/10 partition. This ensures that no firm appears in both the training and evaluation sets.

E.7.3 Tokenization and Loss Masking

Each prompt-completion pair is tokenized using the base model's chat template. We first apply the template to the user message alone to determine the prompt span, then apply it to the full user-assistant pair. Tokens in the prompt span receive loss mask -100; only completion tokens are supervised. When the full sequence exceeds the maximum length (8,192 tokens), we apply left-truncation to retain the tail, ensuring the completion region and its most proximate prompt context are preserved.

E.7.4 QLoRA Configuration

Table A6 reports the default hyperparameters. The base model is loaded in 4-bit quantized form using the NF4 scheme with double quantization and `bfloat16` compute dtype. LoRA adapters are attached to all main attention and MLP projection modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`.

E.7.5 Training and Evaluation Algorithms

Table A6. Default hyperparameters for QLoRA SFT.

Component / Parameter	Default Setting
Maximum sequence length	8,192
Per-device train batch size	1
Per-device eval batch size	1
Gradient accumulation steps	8
Training epochs	2.0
Learning rate	2×10^{-4}
Warmup ratio	0.03
Weight decay	0.0
Learning-rate schedule	Cosine
Logging steps	10
Evaluation / save steps	200
Optimizer	`paged_adamw_8bit`
LoRA rank r	16
LoRA alpha α	32
LoRA dropout	0.05
Target modules	`q,k,v,o,gate,up,down_proj`
Quantization	4-bit NF4 + double quantization
Compute dtype	`bfloat16`

Algorithm 5 Pseudo-label construction and QLoRA SFT

Require: Revised Q2 results with (P_2^{rev}, `q2_answer`); firm-year table with q_{tot}
Ensure: Trained LoRA adapter for calibrated project generation
Merge files on (`gvkey`, `fyear`); drop rows with missing fields
for each retained firm-year **do**
    Parse raw Q2 answer into exactly three project objects
    Extract total investment from the financial JSON in P_2^{rev}
    for each project $i \in \{1, 2, 3\}$ **do**
        Set `IMPLEMENTATION COST`$_i \leftarrow$ total investment (or fallback)
        Set `MARKET VALUE`$_i \leftarrow q_{\text{tot}} \times$ `IMPLEMENTATION COST`$_i$
    end for
    Serialize calibrated projects as completion y_n
end for
Split into train/validation/test using grouped random splits on `gvkey`
Tokenize each (x_n, y_n) with prompt masking and completion-only labels
Load 4-bit base model, attach LoRA adapters, train with $\mathcal{L}_{\text{SFT}}$
Save adapter and tokenizer

Algorithm 6 Held-out evaluation of the SFT adapter

Require: Base model, trained LoRA adapter, held-out JSONL test set
Ensure: Row-level predictions and project-expanded error metrics

Load base model in 4-bit form and attach the trained adapter
for each test prompt **do**
    Generate a completion deterministically (`temperature` = 0, `top_p` = 1)
    Parse the generated JSON array
    if array contains exactly three valid project objects **then**
        Compute $\widehat{q}_1, \widehat{q}_2, \widehat{q}_3$
        Mark `valid_json` = 1
    else
        Mark `valid_json` = 0
    end if
    Write row-level outputs to CSV
end for
Expand valid predictions to project level; compare each $\widehat{q}_i$ with q_{tot}
Report valid-JSON rate, MAE, RMSE, MAPE, and $\text{MAE}_{\log}$

F Example: Varian Medical Systems

We trace the complete pipeline for Varian Medical Systems (VAR, gvkey 11115) in fiscal year 2008, from retrieval through q computation. The example illustrates how the system assembles heterogeneous textual evidence, proposes concrete investment projects, and assigns economic values. It also provides a validity check: we compare the generated projects against Varian's actual subsequent investments, patents, and strategic actions.

Company background. Varian Medical Systems, headquartered in Palo Alto and classified under SIC 3845 (electromedical and electrotherapeutic apparatus), leads the market for radiation oncology equipment and clinical software. Its core products include linear accelerators (LINACs) for cancer treatment, radiosurgery systems, proton therapy equipment, brachytherapy systems, and data management software for radiation oncology clinics. The company operates across three segments and has historically expanded its product portfolio through both internal R&D and targeted acquisitions. Siemens Healthineers acquired Varian in 2021 for $16.4 billion, underscoring the long-run value of the firm's technology platform.

Fiscal year 2008 is a particularly informative case because the firm sits at a strategic inflection point. Varian had completed the $30 million acquisition of ACCEL Instruments GmbH, a proton therapy supplier, in January 2007. It was actively rolling out image-guided radiation therapy (IGRT) products and investing in a nascent security and cargo screening business. These concurrent investment threads—spanning acquisition-driven expansion, organic R&D, and market diversification—test the pipeline's ability to identify qualitatively distinct marginal projects from textual evidence alone.

F.1 Retrieved and Compressed Evidence

For the firm-year (Varian, 2008), the retrieval stage assembles evidence from all five corpora subject to the no-look-ahead constraint: only documents dated before fiscal year 2008 are eligible. The sequential retrieval proceeds as described in Section 3.1, with each stage conditioning on evidence already collected. After retrieval, the compression layer (Section 3.2) rewrites each source-specific block into a structured summary. The resulting evidence bundle is reproduced below.

10-K filings (Items 1 and 7). The firm's own disclosures establish its strategic position:

- Manufactures and markets linear accelerators, simulators for planning cancer treatments, brachytherapy systems, and data management systems for radiation oncology centers.
- Operates in three segments with products serving the Health Care Systems business.
- Focuses on applying advanced technology and developing new products.
- Expanding product offerings through acquisitions, including a joint venture with Mitsubishi Electric Corp. for medical linear accelerators.

Peer firm 10-Ks (same SIC industry). Peer disclosures reveal the competitive environment:

- The industry is highly competitive and changes rapidly, with worldwide patent protection considered important for competing effectively.
- Firms commit significant resources to research and new product development.
- It is difficult to identify specific competitors or estimate the size of primary markets due to the breadth of potential applications.

Patent filings. The firm's innovation frontier:

- Patents describe technologies related to plasma generation, x-ray tubes, microfluidic devices, image data processing, and fluid handling systems.
- Technologies include methods for generating ion beams, preventing filament fusion in electron emitters, and analyzing samples using carrier gases.

Earnings conference calls. Near-term investment signals:

- Acquisition of ACCEL Instruments to supply proton therapy systems ($30 million).
- Developing and rolling out new products, including IGRT and oncology information systems.
- Investing in industrial applications, particularly cargo screening technology (sales slightly over $10 million in the prior year).

- Increasing R&D spending related to IGRT and oncology, focusing on onboard imaging and other IGRT components.

Wall Street Journal articles. Macroeconomic and operational context:

- The company manufactures products formerly purchased from Japan and Korea, potentially providing a cost advantage.
- The company aims to serve blue-chip clients and operates with minimal bureaucracy and a focus on productivity.

The evidence bundle illustrates the complementarity across sources that motivates the multi-source retrieval design. The firm's own 10-K filings establish its core business and acquisition strategy but contain little about specific near-term investment priorities. Earnings calls fill this gap, surfacing three concrete investment threads—proton therapy, IGRT, and cargo screening—that the 10-K treats only at a high level. Patent filings reveal the underlying technological capabilities (ion beams, image data processing) that connect to these directions. Peer disclosures contextualize the competitive intensity and the importance of patent protection, while WSJ articles provide operational context about manufacturing advantages. The pipeline's value lies in assembling these complementary signals; no single source suffices.

F.2 Generated Projects

Given the compressed evidence and a structured financial snapshot from Compustat–CRSP (including total assets, capital expenditures, R&D spending, market capitalization, revenue, and profitability for fiscal year 2008), the generation stage proposes three plausible next investment projects for Varian. Each project includes a description, reasoning grounded in the retrieved evidence, an estimated project-level q, and classification labels.

Project 1: Proton Therapy System Expansion and Optimization.

> *Description.* Expand the production and service capabilities for the ACCEL Instruments proton therapy systems acquired in 2007. Focus on reducing installation times and improving system reliability, addressing revenue recognition concerns highlighted in the conference calls.

Reasoning. The conference calls indicate strong interest in proton therapy. Leveraging the recent acquisition and addressing installation and acceptance issues will capitalize on this demand. The market is growing, and Varian already has a foothold. This project aligns with the company's focus on oncology and whole-system delivery.

Estimated q: 3.50. *Labels:* physical, buy, low compatibility.

Project 2: Enhanced Image-Guided Radiation Therapy (IGRT) Software Suite.

Description. Develop and market a more comprehensive software suite for IGRT, focusing on automated treatment planning and real-time adaptive therapy. Build upon existing onboard imaging technology and address the need for improved efficiency in radiation delivery.

Reasoning. IGRT is a key area of investment identified in the conference calls. Improving the software component offers high margins and leverages existing hardware investments. The patent filings demonstrate ongoing innovation in image data processing. This project addresses the stated need for advanced technology and new product development.

Estimated q: 3.73. *Labels:* intangible, R&D, build, high compatibility.

Project 3: Expansion of Security and Inspection Systems—Cargo Screening.

Description. Increase investment in the cargo screening technology business, focusing on expanding sales internationally and developing new applications for existing component technologies. Target ports and logistics hubs.

Reasoning. The company has a small but growing security business (sales slightly over $10 million). Expanding this segment diversifies revenue streams and leverages existing technology. The WSJ articles highlight the company's ability to manufacture products formerly sourced from Japan and Korea, potentially providing a cost advantage. This aligns with the company's focus on providing whole systems.

Estimated q: 3.88. *Labels:* intangible, build, high compatibility.

Table A7 summarizes the three generated projects and their key attributes. The generated output exhibits several noteworthy features. First, the projects span qualitatively different investment types: acquisition-driven physical expansion (Project 1), organic R&D (Project 2), and market diversification leveraging existing technology (Project 3). This heterogeneity illustrates the extensive margin

of investment—the method captures distinct types of marginal projects rather than producing three scaled-down versions of existing operations. Second, each project's reasoning traces back to specific retrieved evidence: the ACCEL acquisition and conference call signals for Project 1, patent filings in image data processing and IGRT rollout discussions for Project 2, and the cargo screening revenue figure combined with WSJ manufacturing context for Project 3. Third, the projects receive distinct labels along the dimensions used in the main empirical analysis: buy versus build, physical versus intangible, and high versus low compatibility with existing operations. Project 1 is classified as low compatibility (proton therapy via acquisition represents a new capability), while Projects 2 and 3 are classified as high compatibility (software development and technology extension build on existing operations).

Table A7. Summary of Generated Projects: Varian Medical Systems, FY 2008

Project	q	Type	Labels	Primary Evidence Source
1. Proton Therapy Expansion	3.50	Physical	Buy, Low compat.	Earnings calls (ACCEL acquisition, proton therapy demand)
2. IGRT Software Suite	3.73	Intangible, R&D	Build, High compat.	Patents (image processing), Earnings calls (IGRT rollout)
3. Cargo Screening	3.88	Intangible	Build, High compat.	Earnings calls ($10M revenue), WSJ (manufacturing advantage)

Notes. This table summarizes the three investment projects generated by the pipeline for Varian Medical Systems in fiscal year 2008. Each project's estimated q is the ratio of market value to implementation cost. Labels classify projects along dimensions used in the main empirical analysis: physical versus intangible, buy versus build, and high versus low compatibility with existing operations. The primary evidence source column identifies the corpus from which the key signals supporting each project originate.

F.3 q Computation and the Average–Marginal Wedge

The project-level q values are 3.50, 3.73, and 3.88 for Projects 1 through 3, respectively, yielding a firm-year average $q^{AI} = 3.70$.[1] All three projects have $q > 1$, indicating that each creates value and should, under standard investment theory, be undertaken.

We compare q^{AI} against the conventional proxies for this firm-year. Standard average Q—

[1] After supervised fine-tuning calibration (Section 3.5), all three projects converge to $q^{AI} = 3.76$, reflecting the firm-year anchor of $q_{\text{tot}} = 4.00$ from Peters and Taylor (2017). The pre-SFT values reported here preserve cross-project variation in the direct RAG output.

defined as market capitalization divided by the book value of physical capital—is 12.86, while the intangible-adjusted total Q of Peters and Taylor (2017) is 4.00. The wedge $Q^{std} - q^{AI} = 9.1$ is striking: standard average Q exceeds the bottom-up marginal measure by a factor of roughly 3.4×.

Under Hayashi (1982), marginal q equals average Q if the firm exhibits constant returns to scale, operates in perfectly competitive markets, and faces adjustment costs homogeneous of degree one in capital. Varian's large wedge reflects multiple departures from these conditions. First, omitted intangible capital—R&D, patents, organizational know-how—inflates Q^{std} because the denominator captures only physical capital. The intangible-adjusted $Q^{tot} = 4.00$ narrows the gap substantially. Second, Varian earns rents from its dominant market position in radiation oncology, which capitalize into firm value but do not raise the marginal return to new investment. Third, Varian's strategic inflection point in 2008—simultaneously pursuing an acquisition-based expansion into proton therapy (a new capability with low compatibility) and organic R&D in IGRT (a high-compatibility extension)—creates project-level heterogeneity in marginal returns that average measures cannot capture. This pattern illustrates the extensive-margin mechanism of Jovanovic and Rousseau (2014): when the marginal project involves reorienting the firm's capital (Project 1), the investment-q elasticity may be lower than when the project extends existing operations (Projects 2 and 3), consistent with the heterogeneity results in Table 7.

Even the intangible-adjusted Q^{tot} exceeds q^{AI} by 0.24 (approximately 6%), consistent with residual rents not captured by the intangible adjustment. These patterns align with the wedge decomposition in Table 9 and Figure 7, where intangible intensity and profit margins are the dominant drivers of the average–marginal divergence in the healthcare industry.

F.4 Ex-Post Validation

The generated projects were constructed using only information available before fiscal year 2008. We now assess whether the investment directions the pipeline identified align with Varian's actual subsequent trajectory. A clarification is in order: these three projects were not unknown to industry observers—proton therapy, IGRT, and cargo screening were all visible strategic priorities as of 2007. The test is whether the pipeline correctly identifies and prioritizes these directions from heterogeneous textual evidence, assigns them economically meaningful structure (distinct project

types, compatibility labels, and q values), and does so through a systematic, scalable process applied uniformly across thousands of firm-years.

Project 1: Proton therapy expansion. The pipeline identified the 2007 ACCEL acquisition as the basis for a proton therapy expansion project (Figure A1). This direction was actively pursued. In 2008, Varian filed a patent titled "Treatment booth for radiation therapy" (US20090202045A1), describing a radiation system with a curved frame for proton beam delivery (Figure A2). The patent, filed by inventors including then-CEO Timothy Guertin, confirms R&D investment in proton therapy infrastructure immediately following the ACCEL acquisition. Varian subsequently became a leading global supplier of proton therapy systems, though the full buildout of this business occurred over the following decade—a timeline partly influenced by the capital constraints hospitals faced during the 2008–2009 financial crisis.

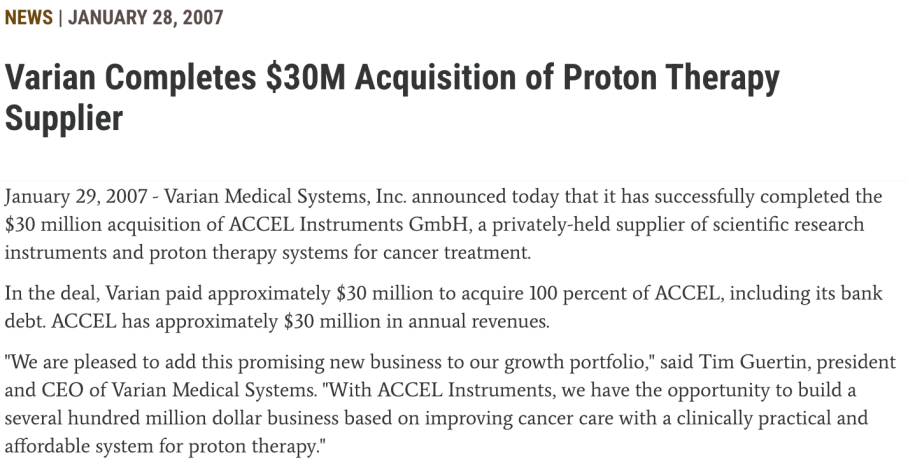
Project 2: IGRT software development. The pipeline identified enhanced IGRT software as a high-compatibility R&D project, grounded in patent filings on image data processing and conference call discussions of IGRT rollout. Varian's subsequent R&D trajectory confirms sustained investment in this area. A 2014 patent on tumor visibility evaluation for IGRT (US10092251B2) describes methods for adaptive therapy using treatment-time imaging (Figure A3), consistent with the generated project's focus on automated treatment planning and real-time adaptive therapy. IGRT was already an active R&D area in 2007; the pipeline's contribution is to identify it as the most plausible high-compatibility intangible investment and to distinguish it from the other two project types.

Project 3: Cargo screening expansion. The pipeline identified cargo screening as a market diversification opportunity, noting approximately $10 million in existing revenue and a domestic manufacturing cost advantage from conference calls and WSJ articles. The security and inspection segment continued to grow in subsequent years, consistent with this identification. We note that the evidence for this project is the thinnest of the three, as the cargo screening business remained a small share of Varian's revenue throughout this period.

In summary, the three generated projects correspond to investment directions that Varian pursued

in the years following fiscal year 2008. The pipeline identified these directions from pre-2008 textual evidence—filings, earnings calls, patents, peer disclosures, and news articles—without access to future information. The value of this example is not that the pipeline "predicted" investments no analyst could foresee, but rather that it systematically extracted, structured, and valued the marginal investment opportunities embedded in heterogeneous text, producing project-level economic objects (with types, compatibility labels, and q values) that accounting-based measures of average Q cannot deliver.

Figures for the Varian Example



NEWS | JANUARY 28, 2007

Varian Completes $30M Acquisition of Proton Therapy Supplier

January 29, 2007 - Varian Medical Systems, Inc. announced today that it has successfully completed the $30 million acquisition of ACCEL Instruments GmbH, a privately-held supplier of scientific research instruments and proton therapy systems for cancer treatment.

In the deal, Varian paid approximately $30 million to acquire 100 percent of ACCEL, including its bank debt. ACCEL has approximately $30 million in annual revenues.

"We are pleased to add this promising new business to our growth portfolio," said Tim Guertin, president and CEO of Varian Medical Systems. "With ACCEL Instruments, we have the opportunity to build a several hundred million dollar business based on improving cancer care with a clinically practical and affordable system for proton therapy."

Figure A1. Varian's January 2007 Press Release: ACCEL Instruments Acquisition

Notes. Press release announcing Varian's $30 million acquisition of ACCEL Instruments GmbH, a proton therapy supplier. This acquisition, identified by the retrieval system through earnings call transcripts, forms the evidential basis for Generated Project 1 (Proton Therapy Expansion).

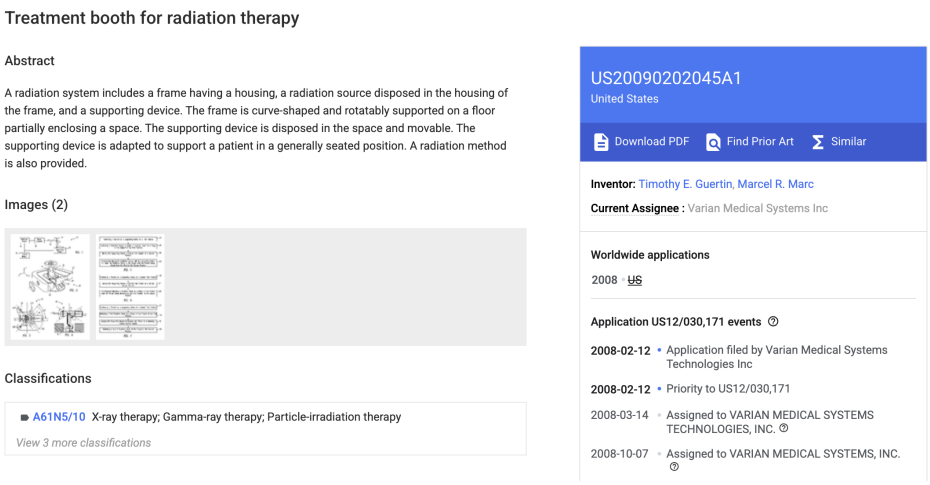


Figure A2. Varian Patent Filing (2008): Treatment Booth for Radiation Therapy

Notes. Patent US20090202045A1 filed by Varian Medical Systems in 2008, describing a curved-frame radiation system for proton beam delivery. The filing confirms active investment in proton therapy R&D following the ACCEL acquisition, consistent with Generated Project 1.

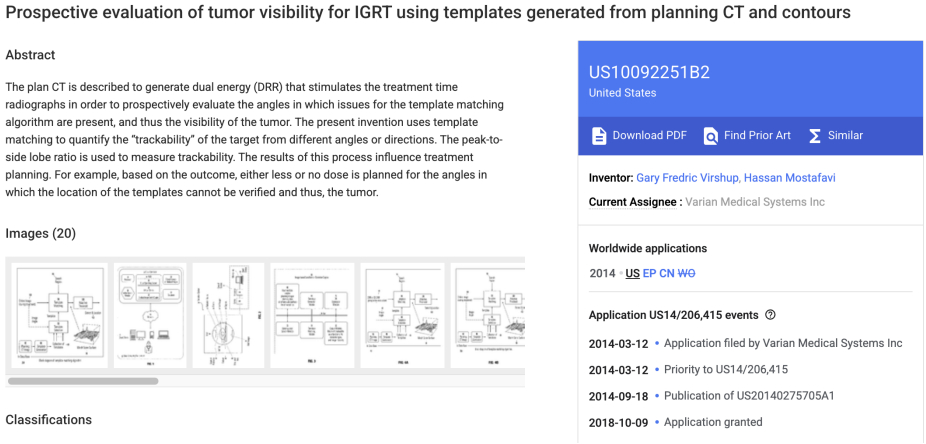


Figure A3. Varian Patent Filing (2014): Tumor Visibility Evaluation for IGRT

Notes. Patent US10092251B2 filed by Varian Medical Systems in 2014, describing methods for evaluating tumor trackability during image-guided radiation therapy. The patent on adaptive therapy technology confirms sustained R&D investment in IGRT, the direction identified as the primary intangible investment opportunity by the pipeline in 2008 (Generated Project 2).



CLINICAL NEWS | RADIATION ONCOLOGY/THERAPY

Varian wins U.K. contract, unveils cargo screening technology

AuntMinnie.com staff writers
May 25, 2005

Radiation therapy vendor Varian Medical Systems reported that it has been selected as a supplier of linear accelerators in the British government's capital investment program for upgrading cancer treatments in its hospitals.

In other news, Varian said that it has developed a new type of x-ray linear accelerator that will enable cargo screening systems to automatically alert operators if suspicious materials are detected.

Figure A4. News Report: Varian Cargo Screening Technology

Notes. News article (May 2005) reporting that Varian won a U.K. contract for linear accelerators and unveiled new cargo screening technology. This article, which predates the prediction year, illustrates the type of evidence available to the retrieval system when identifying cargo screening as an investment opportunity (Generated Project 3).