# Beyond GICS: Vector Search over Business Descriptions to Discover Data-Driven Company Comparables

Quinn Lambert

## Abstract

Finding a company's peer set is a routine, yet very causative, task: relative-value trades, pairs construction, index and ETF design, and multiples-based valuation all derive significant relevancy by deciding which firms are comparable. The market's default answer is GICS, a committee-assigned taxonomy keyed to a firm's primary source of revenue and earnings. I test an alternative that defines peers by what a firm actually does. I embedded the business descriptions of 502 S&P 500 firms with a sentence-transformer model, loaded the vectors into a Qdrant vector database, and queried each firm's ten nearest semantic neighbors. The neighbors land in the firm's own GICS sector 65.2% of the time against a size-based random expectation of 10.9% (6x) so the embeddings recover sector structure strongly. The remaining third of crossings is not noise. It concentrates at a few specific, economically sensible sector boundaries; it runs asymmetrically, from small heterogeneous sectors into large coherent ones; and it isolates individual firms (Berkshire Hathaway, Broadridge) that GICS arguably misplaces. And the natural next test is whether embedding-defined peer baskets co-move more tightly than GICS baskets.

## 1. Introduction

The first of many questions behind a great deal of tedious desk work is understanding "who are this company's peers?", and the answer is rarely questioned. It sits underneath relative-value and pairs trades, where you need a comparable to trade a name against; multiples-based valuation, where you price a firm off the multiples of its comps; ETF and index construction, where a sector fund is a peer basket by definition; and factor and risk models, where sector exposure is a standard control. A wrong peer set propagates into every one of these, and so the mapping of firm to comparables is integrally load-bearing.

The default peer map is the Global Industry Classification Standard. MSCI and S&P assign each firm to one of eleven sectors, then to finer industry and sub-industry buckets, based primarily on its main source of revenue and earnings. GICS is stable, widely adopted, and easy to reason about, which is exactly why it is the default. But these same properties also happen to be its biggest limitation. It is a static, single-label taxonomy maintained by committee, so it can strand a conglomerate with several real businesses in one box, and it can lag a firm whose actual operations have drifted away from the line it was first filed under.

So that raises the question as to what other starting points might be superior, such as by defining peers, quite blankly, by what a firm asserts it supposedly does. Every constituent publishes a business description, and if two firms describe their operations in similar language, they are doing similar things regardless of which revenue line a committee keyed on. I turn each description into a vector with a sentence-embedding model and treat the nearest vectors as a firm's data-driven peers.

Nearest-neighbor search over dense embeddings is the workload a vector store is built for, and it is the natural non-relational tool for this job. There is no fixed schema to join on, the query is simply "find the most similar rows," and similarity is a geometric operation rather than a SQL predicate. That motivates the NoSQL choice at the center of the project: I use Qdrant rather than forcing approximate similarity into a relational table.

This sets up a clear research question: do embeddings of business descriptions recover GICS sector structure, and where they diverge from it, do they reveal economically meaningful peer relationships that GICS misses? My main hypothesis (H1) is that nearest semantic neighbors mostly stay in-sector as the embeddings encode real structure, but a non-trivial share crosses sectors, and those crossings are systematic rather than random. Three sub-hypotheses explicate this further. H1a: sector cohesion varies, but because raw self-retention conflates cohesion with sector size, it must be judged against a size-based baseline; homogeneous-vocabulary sectors should stay far above chance, while large heterogeneous sectors should sit lowest relative to chance. H1b: specific firms (conglomerates and mislabeled fintech, for instance) are placed by GICS away from their description peers. H1c: the crossings are directional, with small heterogeneous sectors reaching into large coherent ones in a sort of donor-to-attractor pattern.

The rest of the report basically describes the data and pipeline, then works through four results—overall recovery, the structure of the crossings, firm-level case studies, and size-adjusted cohesion—before stating conclusions and the co-movement test that follows from them.

## 2. Methods

### 2.1 Data sources

Ground-truth labels and the firm universe come from the Wikipedia list of S&P 500 constituents, which supplies each firm's ticker and its GICS sector and sub-industry. The business text comes from the longBusinessSummary field returned by yfinance. Of the listed constituents, one failed to return a usable summary and was dropped, leaving N = 502 firms spanning all eleven GICS sectors; the S&P 500 membership snapshot was taken in June 2026. One limitation belongs here, stated plainly: the longBusinessSummary is a short paragraph, not the full Item 1 of a 10-K, so the embedding sees a compressed view of each business. I also collected each firm's GICS sub-industry; I do not use it in the analysis below to avoid scope creep, but it is available for a finer-grained follow-up if desired.

### 2.2 Embedding

I embedded each description with all-MiniLM-L6-v2, a 384-dimensional sentence-transformer. This choice was made deliberately as it runs locally with no API dependency, it is a standard general-purpose sentence encoder, and 384 dimensions are cheap to store and query. Vectors are cosine-normalized, so nearest-neighbor search by inner product is equivalent to cosine similarity.

### 2.3 Vector store

I loaded the 502 vectors into a Qdrant collection running in Docker, configured for 384-dimensional vectors with cosine distance. Each point carries a payload with the firm's ticker, name, GICS sector, and sub-industry, so a similarity query returns labeled neighbors directly. I confirmed the populated collection in the Qdrant dashboard before running any queries. Two configuration details matter for

what follows. Distance is cosine over the 384-dimensional vectors, which matches the cosine-normalized embeddings so that a similarity score is just their dot product. And because the 502-vector collection sits below Qdrant's full-scan threshold, retrieval is exact rather than approximate HNSW, so the neighbor rates reported below carry no approximate-nearest-neighbor error; at larger scale, Qdrant's HNSW index would trade exactness for speed. The sector payload also makes the collection queryable rather than only searchable—a k-NN call can be filtered on payload, for instance to constrain neighbors to a chosen sector, or to exclude a firm's own sector and surface only the cross-sector crossings.

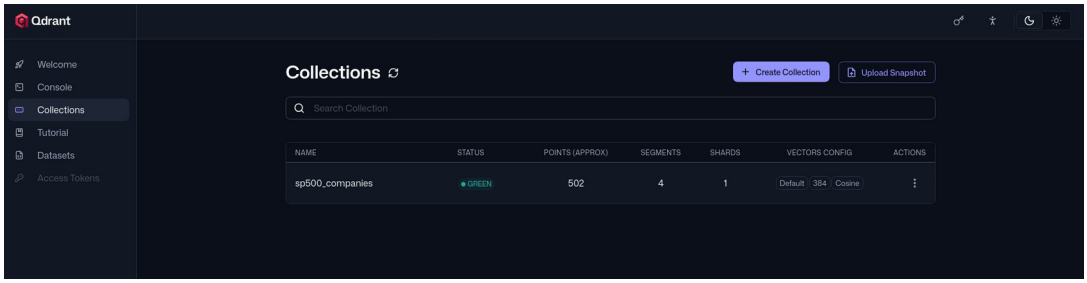


***Figure A.*** *The populated Qdrant collection (502 points, 384-dim, cosine), confirming the live vector store.*

## 2.4 Analysis definitions

For each firm I retrieved its k = 10 nearest neighbors, excluding the firm itself. A firm's cross-sector neighbor rate is the fraction of those ten neighbors whose GICS sector differs from its own; the intra-sector rate is the complement. Aggregated to the sector level, I build a row-normalized affinity matrix whose (i, j) entry is the share of sector i's neighbors that land in sector j; in Section 2.5 I convert these raw shares into a size-adjusted lift, so that attraction is measured net of how large the receiving sector happens to be. I flag a firm as a likely mismatch when its cross-sector rate is at least 0.6 and its most common neighbor sector is not its own GICS sector.

## 2.5 Baselines

To judge any of these rates I need to know what they would be by chance. If neighbors were drawn at random, the probability that a given neighbor shares a query firm's sector is just the share of other firms in that sector. Pooled across the index, the expected intra-sector rate is 10.9%. Because sectors differ enormously in size, I also compute a per-sector chance term and report each sector's observed self-retention as a lift over its own baseline. This is what separates genuine cohesion from the mechanical advantage that large sectors have in finding their own kind.

```
# global expected intra-sector rate under random neighbors
N = sum(counts.values())
exp_intra = sum(n*(n-1) for n in counts.values()) / (N*(N-1))   # = 0.109

# per-sector chance and lift (used for the cohesion table / Figure 4)
chance_s = (counts[s]-1)/(N-1)        # expected intra for a query in sector s
lift_s   = obs_intra_s / chance_s     # obs_intra_s = 100 - porosity_s

# generalize the same baseline to every (i, j) cell, not just the diagonal
exp_ij  = (counts[j] - (i == j)) / (N - 1)   # expected share if neighbors
were random
lift_ij = obs_ij / exp_ij                    # >1 attraction, <1 avoidance, 1
= chance
# 2,000-iteration label-permutation null gives a per-cell p-value
```

The same baseline generalizes from the diagonal to every cell of the affinity matrix. For sector i pointing to sector j, the expected share of i's neighbors landing in j, if neighbors were drawn at random, is (count_j − [i = j]) / (N − 1)—the −1 on the diagonal simply excludes a firm from being its own neighbor—and the cell's lift is its observed share divided by that expected share. This is exactly the chance term already used for self-retention in the cohesion analysis, now applied to the off-diagonal cells as well, so it divides out both the sending and the receiving sector's size and removes the receiving-sector size confound that a raw row percentage leaves in. A lift of 1.0 is chance, above 1 is genuine attraction, and below 1 is avoidance. To separate real structure from sampling noise I attach a significance test: for each cell I compare its observed share against a null built by randomly permuting the GICS labels 2,000 times and recomputing the matrix, which yields a per-cell p-value. This is a standard observed-over-expected enrichment ratio—lift in association analysis, or a configuration-model null in network terms—now applied cell by cell rather than only to self-retention.

### 2.6 Reproducibility

Stochastic steps, including the UMAP projection, were run with a fixed random_state. The pipeline uses all-MiniLM-L6-v2 via sentence-transformers for embeddings, Qdrant for storage and retrieval, and standard scientific-Python libraries for the affinity and baseline computations. Figure 1 colors points by their eleven GICS sectors; note that two of the categorical colors (Communication Services and Utilities) are close, so that figure is best read for cluster structure rather than precise sector-by-color identification.

# 3. Analysis

## 3.1 Overall: the embeddings recover GICS structure

The headline result is that the embeddings recover GICS structure strongly, but not perfectly; and it's the imperfection which is actually interesting. Across all 502 firms, 65.2% of nearest neighbors fall in the firm's own GICS sector, against the 10.9% expected if neighbors were random. That is a 6.0x lift over chance. The bracketing logic matters here: a model that merely reproduced GICS would show a near-zero cross-sector rate and a purely diagonal affinity matrix, while pure noise would show a cross-sector rate near the 89.1% random ceiling and a uniform matrix. The observed 34.8% cross-sector rate sits much closer to the informative end, so whatever the model absorbed from generic web text, it encodes sector membership strongly.

Figure 1 shows this directly. A UMAP projection of the 502 embeddings, colored by GICS sector, separates into visible sector neighborhoods: Financials anchors one corner, Health Care another, and Utilities and Energy hold their own regions on the lower right. There is one caveat to consider before leaning on the picture: UMAP is non-metric, so only local neighborhoods are trustworthy. I use it to show that clusters form, not to read global distances off the page. The points that sit inside a foreign-colored region are a visual preview of the firm-level mismatches in Section 3.3.

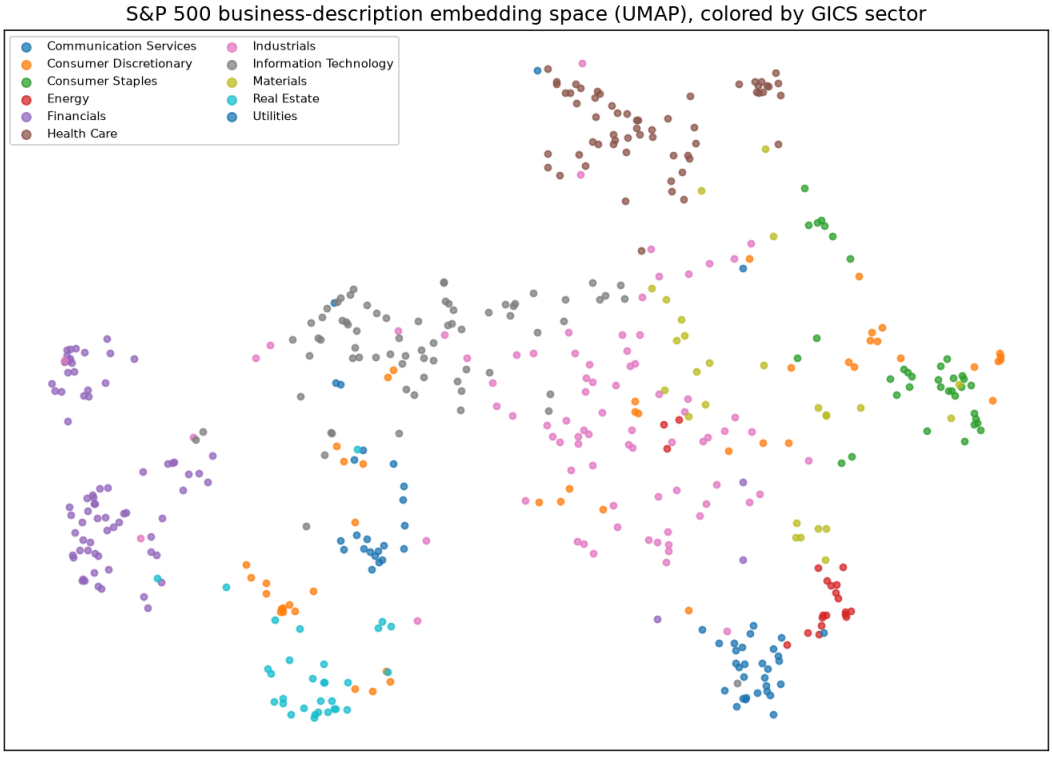


***Figure 1.*** *UMAP projection of the 502 S&P 500 business-description embeddings (all-MiniLM-L6-v2, 384-dim), colored by GICS sector. Sectors form visible neighborhoods; UMAP is non-metric, so only local structure is interpretable.*

## 3.2 The crossings are systematic and asymmetric

The cross-sector neighbors are not scattered evenly. They pile up at a few specific boundaries, and they flow mostly in one direction. Reading attraction straight off the raw row shares is tempting but misleading, though: a plain row percentage rewards whichever receiving sector is biggest, because a

large sector has more firms available to be anyone's neighbor and so captures a larger slice of every row regardless of any real semantic pull. Figure 2 therefore reports each cell as lift over a firm-count baseline (observed share ÷ expected share) rather than as a raw percentage—the same chance baseline used for self-retention in Section 2.5, now applied to every cell, which divides out both the sending and the receiving sector's size. A lift of 1.0 is exactly chance, above 1 is genuine attraction, and below 1 is avoidance. The raw row-normalized shares are still a fair description of where a sector's neighbors land—they simply conflate size with pull, which is why attraction itself is measured as lift.

Measured this way, the off-diagonal structure concentrates in a handful of economically sensible cells. The strongest genuine cross-sector affinities, by lift, are Energy → Utilities (about 3.7×), Consumer Discretionary → Consumer Staples (2.8×), Materials → Consumer Staples (2.5×), Materials → Energy (2.3×), Consumer Discretionary → Real Estate (1.7×), and Materials → Industrials (1.5×)—the chemicals-and-packaging, power-generation-and-pipeline, and shared consumer-retail overlaps you would expect, plus a Communication Services → Information Technology pull (1.4×) that echoes the 2018 GICS reclassification of media and internet names. These are not scattered noise: against a 2,000-iteration label-permutation null, 99 of the 121 cells differ from chance at $p < 0.05$, and the cells that do not are marked with a dot in Figure 2.

The more interesting observation is that these flows are asymmetric, and the asymmetry not only survives the size correction but comes out cleaner. Materials reaches up toward Industrials at a lift of 1.52× ($p = 0.0005$)—a real, statistically significant attraction—while the reverse direction, Industrials → Materials, sits at 0.67× ($p = 0.91$), statistically indistinguishable from chance. Industrials does not reach back: there is simply no pull in that direction, which is different from active avoidance. In raw landing shares the contrast looked like roughly 6.9× (24% versus 3.5%); once sector size is divided out it is about 2.2× (1.52× versus 0.67×), and it is now significance-tested rather than asserted. The pattern is still a *donor-to-attractor* structure: small, heterogeneous sectors (Materials, Energy, Communication Services, Consumer Discretionary) reach into large, coherent ones (Industrials, Information Technology, Utilities, Consumer Staples), and that does not follow to the same degree vice versa. The mechanism is pretty intuitive. Effectively, a coherent sector defines a tight, dense region of description-space, while a diffuse sector's firms scatter, and many of them land in whichever dense region they sit closest to.

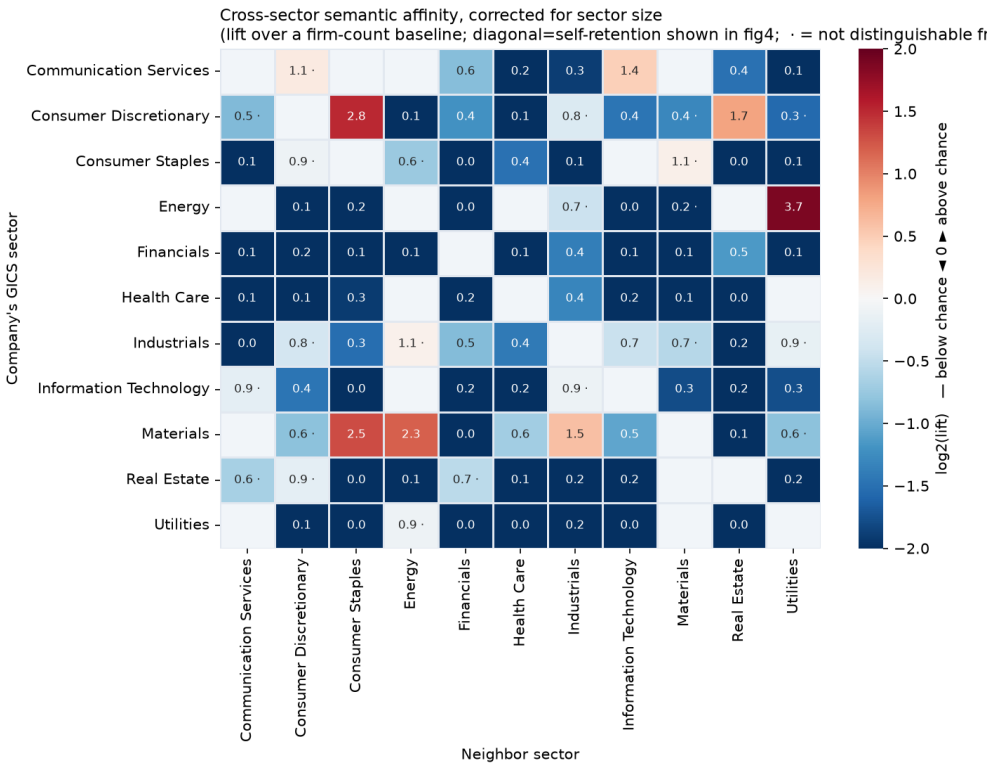


***Figure 2.*** *Cross-sector semantic affinity as lift over a firm-count baseline (observed ÷ expected share; red = above chance, blue = below chance, on a $\log_2$ scale centered at chance). Diagonal self-retention is shown separately in Figure 4; · marks cells not distinguishable from chance at $p < 0.05$ (2,000-iteration label-permutation null).*

### 3.3 Case studies: where GICS is the outlier

At the firm level, the crossings isolate names where the embedding arguably has it right and GICS has it wrong—though it's true that the confidence varies by case. Table 1 summarizes four, and Figure 3 lists the actual nearest neighbors and cosine similarities for three of them. Broadridge appears to be the cleanest miss: GICS files it under Industrials, but all five of its nearest neighbors are banks and exchanges (State Street, Citigroup, Raymond James, BNY Mellon, and Cboe) which are at cosines of 0.57 to 0.59. It is financial-infrastructure fintech that happens to sit in the Industrials bucket, and the embedding has no hesitation about where it belongs.

Berkshire Hathaway is a bit of a different and more transparent case. GICS calls it Financials, classifying off its insurance float, but its neighbors span Industrials (Norfolk Southern, Union Pacific, Huntington Ingalls), Energy (Baker Hughes), and Consumer Discretionary (Tractor Supply) at lower cosines of 0.47 to 0.50. Here the embedding is not so much disagreeing with GICS as revealing that Berkshire has no clean peer set at all: it is just fundamentally a genuinely multi-sector conglomerate acting simply as a Buffet-derived vessel for immense wealth that any single label distorts. Ecolab (Materials by GICS, but four of its five nearest neighbors are lab and diagnostics names—Agilent, Waters, Danaher, and Idexx—with Clorox the lone consumer-staples neighbor, a sensible peer for a hygiene and infection-prevention business) and Aptiv (Consumer Discretionary on paper, software-defined-vehicle technology in the description) round out the pattern. The throughline is consistent: GICS classifies by where the earnings come from, the embedding by what the business

does, and the two part ways exactly where a firm's earnings engine and its operations have drifted apart.

***Table 1.*** *Case-study firms flagged as mismatches (cross-sector rate ≥ 0.6 with a foreign modal neighbor sector). Cosine ranges are taken from Figure 3.*

| Firm | GICS → Semantic | Neighbor cosine | Read |
|---|---|---|---|
| Berkshire Hathaway (BRK.B) | Financials → Industrials | 0.47–0.50 | Actually just multi-sector. Neighbors span Industrials (Norfolk Southern, Union Pacific, Huntington Ingalls), Energy (Baker Hughes), and Consumer Discretionary (Tractor Supply). GICS keys on the insurance float; the embedding reads the operating businesses. No clean single-label peer set exists. |
| Broadridge (BR) | Industrials → Financials | 0.57–0.59 | High-confidence miss. All five nearest neighbors are banks and exchanges (State Street, Citigroup, Raymond James, BNY Mellon, Cboe). Financial-infrastructure fintech filed under Industrials. |
| Ecolab (ECL) | Materials → Health Care | 0.44–0.49 | Water, hygiene, and infection-prevention. Four of five nearest neighbors are lab and diagnostics names (Agilent, Waters, Danaher, Idexx); Clorox is the lone consumer-staples peer, fitting for a hygiene business. |
| Aptiv (APTV) | Consumer Discretionary → Information Technology | — | Auto-components on paper; the description is ADAS and software-defined-vehicle technology. |

BRK.B (GICS: Financials) — nearest semantic neighbors

| Ticker | Name | Sector | Cos |
|---|---|---|---|
| NSC | Norfolk Southern | Industrials | 0.50 |
| BKR | Baker Hughes | Energy | 0.50 |
| UNP | Union Pacific Corporation | Industrials | 0.47 |
| TSCO | Tractor Supply | Consumer Discretionary | 0.47 |
| HII | Huntington Ingalls Industries | Industrials | 0.47 |

BR (GICS: Industrials) — nearest semantic neighbors

| Ticker | Name | Sector | Cos |
|---|---|---|---|
| STT | State Street Corporation | Financials | 0.59 |
| C | Citigroup | Financials | 0.59 |
| RJF | Raymond James Financial | Financials | 0.58 |
| BNY | BNY Mellon | Financials | 0.58 |
| CBOE | Cboe Global Markets | Financials | 0.57 |

ECL (GICS: Materials) — nearest semantic neighbors

| Ticker | Name | Sector | Cos |
|---|---|---|---|
| CLX | Clorox | Consumer Staples | 0.49 |
| A | Agilent Technologies | Health Care | 0.48 |
| WAT | Waters Corporation | Health Care | 0.47 |
| DHR | Danaher Corporation | Health Care | 0.47 |
| IDXX | Idexx Laboratories | Health Care | 0.44 |

***Figure 3.*** *Nearest semantic neighbors, with cosine similarity, for three case-study firms. Broadridge's neighbors are all Financials; Berkshire's span several sectors; Ecolab's are Health Care lab and diagnostics names.*

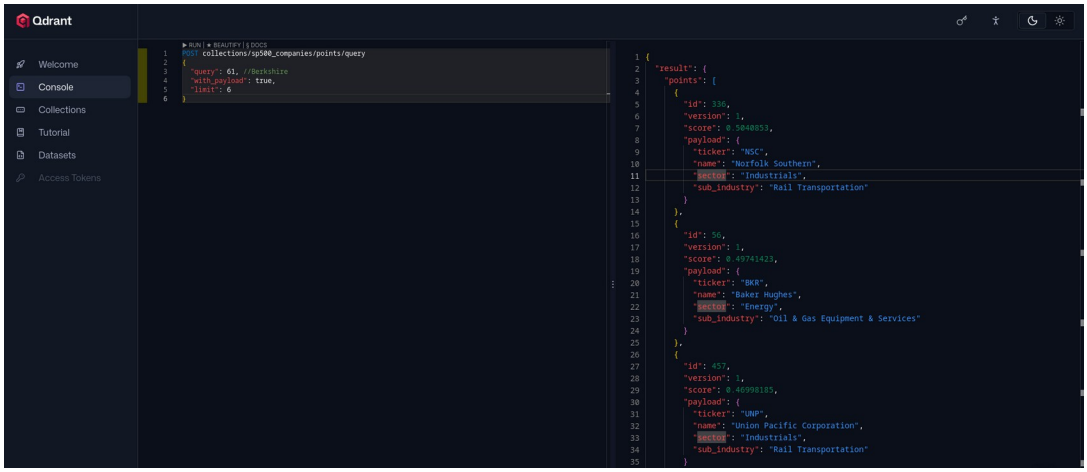


***Figure B.*** *A live nearest-neighbor query in Qdrant using the stored vector of Berkshire Hathaway (point 61); the returned peers are railroads and energy-equipment firms (Industrials and Energy), not Financials.*

## 3.4 Cohesion, once you adjust for sector size

Sectors differ a whole bunch in how tightly they manage to hold together, but you have to adjust for size before ranking them. Doing so overturns the obvious reading. Raw self-retention, in Table 2, runs from Utilities at 91% down to Materials at 25%. Taken at face value, that says Materials barely exists as a coherent sector. But Materials has only 26 firms, so its random baseline is low too, and once observed retention is divided by each sector's own chance term the picture changes.

Figure 4 shows the size-adjusted lift. Energy (15.4x) and Utilities (15.1x) top the ranking; these are tight, homogeneous-vocabulary sectors whose firms describe themselves in nearly interchangeable language. Materials lands mid-pack at 4.9×, not the outlier its raw number implied. The genuinely least cohesive sectors, relative to chance, are Industrials at 3.3× and Consumer Discretionary at 3.8×: large sectors that should have no trouble finding their own kind and still do not, because they are catch-alls spanning unrelated businesses. This is exactly the modulation H1a anticipated. Cohesion tracks vocabulary homogeneity, not headline retention, and the size adjustment is what makes that visible.

***Table 2.*** *Sector cohesion, raw and size-adjusted, sorted by lift. Self-retention is observed intra-sector neighbor rate; chance is the per-sector random baseline (n− 1)/(N− 1); lift is their ratio.*

| Sector | n | Self-retention | Cross-rate | Chance | Lift |
|---|---|---|---|---|---|
| Energy | 21 | 61.4% | 38.6% | 4.0% | **15.4×** |
| Utilities | 31 | 90.6% | 9.4% | 6.0% | **15.1×** |
| Communication Services | 23 | 51.7% | 48.3% | 4.4% | **11.8×** |
| Real Estate | 31 | 70.3% | 29.7% | 6.0% | **11.7×** |
| Consumer Staples | 36 | 75.6% | 24.4% | 7.0% | **10.8×** |
| Health Care | 59 | 84.4% | 15.6% | 11.6% | **7.3×** |
| Financials | 75 | 83.5% | 16.5% | 14.8% | **5.7×** |
| Materials | 26 | 24.6% | 75.4% | 5.0% | **4.9×** |
| Information Technology | 72 | 66.8% | 33.2% | 14.2% | **4.7×** |
| Consumer Discretionary | 48 | 35.6% | 64.4% | 9.4% | **3.8×** |
| Industrials | 80 | 51.6% | 48.4% | 15.8% | **3.3×** |

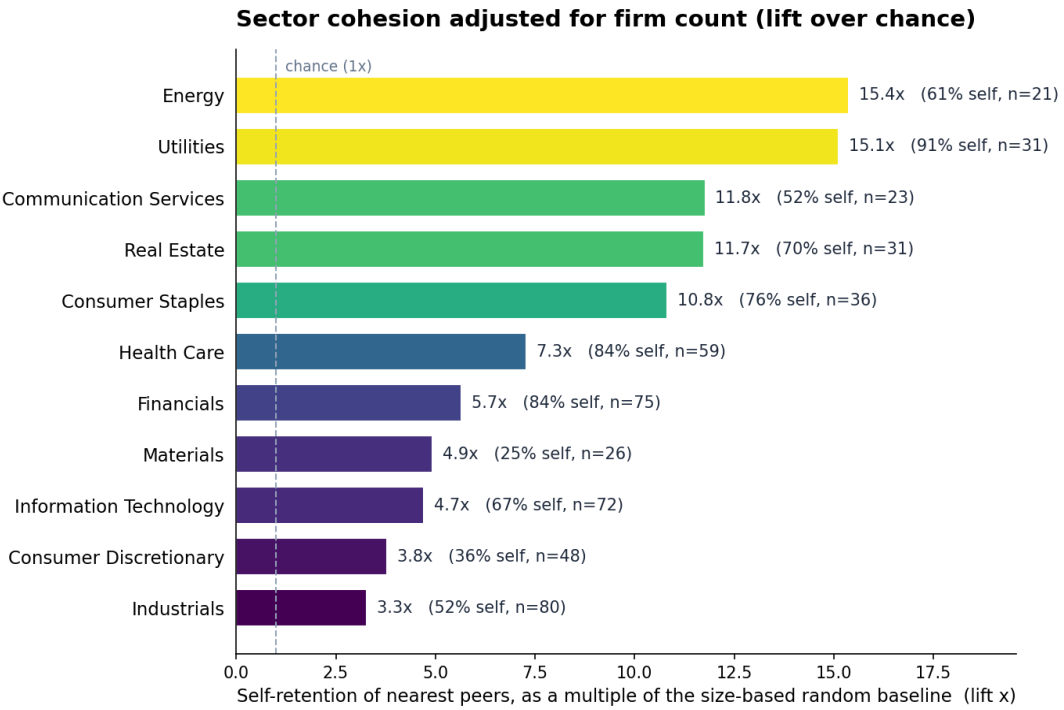


***Figure 4.*** *Sector cohesion as a multiple of the size-based random baseline (lift over chance). Energy and Utilities are tightest once size is accounted for; Industrials and Consumer Discretionary are loosest, despite middling raw self-retention.*

## 4. Conclusions

Embeddings of business descriptions recover GICS sector structure—neighbors stay in-sector roughly six times more often than chance—and the places where they diverge are systematic, asymmetric, and economically meaningful rather than random error. The model agrees with GICS most of the time, and where it disagrees, it tends to disagree for reasons a more involved person would recognize: an evolving operating profile, a conglomerate with no single home, or a label keyed to an earnings engine that no longer describes the business.

That makes the embedding useful in two directions. As a constructive tool, it yields data-driven peer baskets for relative-value, pairs, and risk work that do not depend on a committee's single-label call. As a diagnostic, the mismatch list and the donor-attractor map flag where GICS-based products (sector ETFs and index sleeves) may be carrying firms that do not behave like their labelmates.

The honest limitations are worth stating in one place. The longBusinessSummary is shorter than a full filing, so the model works from a compressed description. Semantic similarity is not the same as return co-movement, which is the test that actually matters for trading. all-MiniLM is a general-purpose model with no finance tuning, the analysis is a single point-in-time snapshot, and both k = 10 and the 0.6 mismatch threshold are choices. UMAP is non-metric and is used only for local structure. One earlier confound is worth flagging precisely because it has been addressed: read as raw shares, both self-retention and the cross-sector affinity matrix conflate cohesion and attraction with sector size, since a large sector mechanically captures more of every neighbor list. Every cell—diagonal and off-diagonal alike—is therefore reported as lift over a firm-count baseline rather than as a raw percentage, which divides out the receiving sector's size as well as the sending sector's.

Controlling for size this way is what separates genuine semantic attraction from a size artifact, and it is what lets the surviving Materials-to-Industrials asymmetry stand as a size-corrected, significance-tested result rather than a possible consequence of Industrials being the largest sector.

The next test writes itself: do embedding-based peer baskets co-move more tightly than GICS baskets? If they do, the semantic peers are not just plausible but they are tradable. Beyond that, the data already supports pushing to GICS sub-industry granularity, tabulating the 103 flagged mismatches by origin sector, swapping in a finance-tuned embedding model and full 10-K text, and tracking how a firm’s semantic neighborhood drifts over time.