Morningstar Equity Comparables Methodology

Morningstar
April 2026
Version 2.0

Contents

Author

Charles Gross
Director of Quantitative Research
Charles.Gross@morningstar.com

Mikhail Avens
Quantitative Analyst
Mikhail.Avens@morningstar.com

Executive Summary

Morningstar developed the Morningstar Equity Comparables system to give investors and financial professionals an objective benchmark for covering companies. Morningstar Equity Comparables is genuinely different from other industry classification schemes. Traditional peer company analysis is either exceedingly expensive and bespoke or relies heavily on hierarchical industry classifications that often fail to capture the nuanced business model similarities crucial for accurate comparable company identification.

This methodology presents a novel large language model-based approach that leverages natural language processing to identify peer companies through automated analysis of business descriptions and operational characteristics.

Validation by our in-house equity analysts gives us confidence that our latest version of Equity Comparables significantly surpasses the quality of traditional algorithmic comparable company offerings. It is also substantially more scalable than leveraging specialists to create and maintain custom sets of comparables, allowing us to deploy this to more than 35,000 companies globally.

Key Takeaways

- LLM-based peer analysis overcomes the rigid limitations of traditional sector-based classifications by analyzing business descriptions directly.
- A multistage validation framework with expert judging ensures high-quality peer selections.
- Automated quality control based on sector alignment and similarity scores optimizes final peer company recommendations.
- The system processes thousands of companies while maintaining strict pure-play focus criteria.
- The process is agnostic of a company's trading status, automatically generating privately held peers for publicly listed companies when relevant.

Thinking Outside the Hierarchy

How should we classify businesses? How do you justify every decision in constructing a taxonomy? How do you avoid catch-all "other" categories? How do you know the classification is correct in some sense?

Our argument is that any taxonomy of business will end up telling us more about the person classifying the objects being classified. We believe the problems with hierarchical classification lead to poorer results when it is employed in valuations or risk management. In this document, we present a better technique, which we call Morningstar Equity Comparables, based on pure business similarity.

It is natural for people to categorize objects in hierarchies or taxonomies–a top-down classification that operates by a mechanism of class inclusion. The General Industry Classification System from MSCI and Standard and Poor's is perhaps the best-known system today, but many similar systems exist, and they are pervasive in the finance industry. We will proceed to criticize GICS because we expect it to be most familiar to the reader, but our criticism applies equally to all top-down taxonomic business classifications, including SIC, NAICS, and Morningstar's own Global Equity Classification Standard. Presently in GICS, the universe of businesses is divided into 10 sectors, 24 industry groups, 68 industries, and 154 subindustries. Once an industry is placed in a subindustry—for example, Peabody Coal common stock might be placed in the coal subindustry—because of the strict hierarchy in the classification, it must also be placed in the successive higher levels of the hierarchy, such as the oil, gas, and consumable fuels industry, the energy industry group, and the energy sector.

But why look at the similarity of businesses anyway? Similar businesses will tend to have similar cash flows, which market prices implicitly discount at a similar rate, which gives a measurement of the comparability of businesses two broad purposes. First, valuation: Given a business, an analyst may wish to understand similar businesses for a comparable business-based relative valuation. Second, risk management: Uncertain future cash flows from similar businesses will tend to be correlated, which will make the market returns of their securities correlated, and many portfolio construction methodologies discourage investing in assets with correlated returns.

Our system can directly capture the similarity of firms in what would be different sectors in hierarchical systems. While we criticize hierarchical systems for their poor performance, they still have the advantage of being simple to understand, and once the hierarchy is defined, it becomes straightforward to determine where an industry is in terms of other industries. By constructing a fragment of a "universal language"[1] to describe businesses, an investor can at least simplify the problem of choice and process the vast amounts of information efficiently.[2] For instance, with GICS, we can infer very simply that coal companies are energy companies, closely related to other oil and gas companies. For indexing purposes—or building an exchange-traded fund—this means that an "Energy Index" has a clear definition. We think hierarchical systems are superior to ours in this indexing and ETF role.

But these bright boundaries between groups and subgroups are any hierarchical system's Achilles heel. The difficulty with this particular example of coal—and we argue for many examples—is that these categorical judgements become a problem around the boundaries. The world of business is not so simple as to be easily squeezed into a hierarchy. For instance, there are two uses of coal: coking coal and thermal coal. Coking coal is used for steelmaking (materials in GICS), thermal coal for electricity and heating (energy in GICS). Furthermore, coal is mined, which suggests the same cost structure as a

1 The Analytical Language of John Wilkins, Jorge Luis Borges.

2 Barberis, N., & Shleifer, A. (2003). "Style investing." Journal of Financial Economics, Vol. 68, P.162

mining company. So, the classification of Peabody Coal as an energy stock is a bit arbitrary. This hierarchy suggests that it's equally wise to diversify the risk in coal with steel businesses as with healthcare businesses: It can easily be seen how naïve application of the classification system might lead to poor investment choices. Every hierarchical classification system with clear-cut boundaries will be weak around these kinds of corner cases. We don't suggest that GICS is wrong to put coal in the energy sector. We argue instead that employing a hierarchical taxonomy leads inevitably to boundary cases where the industry classification becomes arbitrary, so for a material portion of the subclasses or businesses, there is no clear final place where the subclass or a business should sit within the hierarchy.

Morningstar Equity Comparables Overview

Morningstar Equity Comparables is an automated system that leverages the natural language understanding capabilities of large language models to identify peer companies based on detailed business descriptions rather than predetermined taxonomies. By analyzing how companies describe their own operations, the system can identify peers that traditional classification schemes might miss while excluding companies that appear similar at a high level but operate with fundamentally different business models.

The approach is particularly valuable for:

- Pure-play analysis: Identifying companies where the target industry represents their primary business focus.
- Cross-sector comparability: Finding operational similarities that transcend traditional industry boundaries.
- Dynamic peer selection: Adapting to evolving business models and emerging industries.

A New LLM-Based Approach

Our methodology employs a sophisticated processing pipeline that analyzes business filings to identify peer companies based on operational similarity rather than predetermined classification.

Core Methodology Principles

The system operates on several key assumptions that leverage the fundamental architecture and mathematical properties of LLMs, specifically implemented through our retrieval-augmented generation framework.

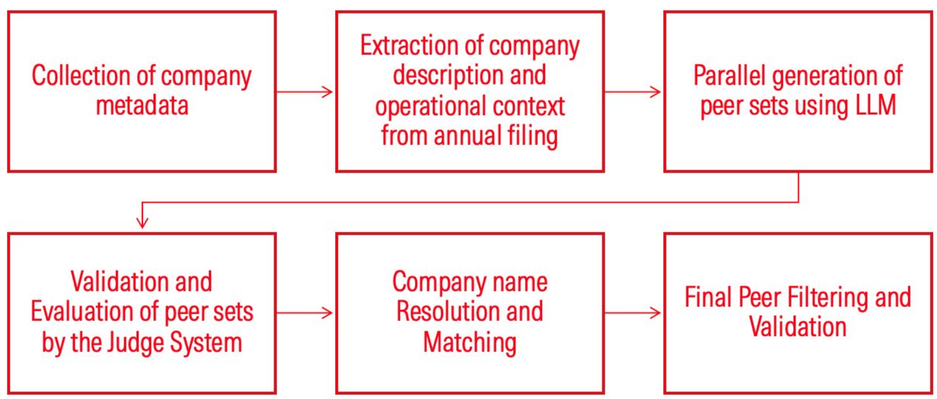
- Business Description Validity: Companies that operate similar business models will describe their operations using similar terminology and concepts

- Primary Business Focus: True peer companies must derive a plurality (>30%) of their revenue from the same primary industry
- Operational Alignment: Beyond industry focus, peers should demonstrate similar business characteristics, such as sales models and operational structures. The LLM identifies these nuanced similarities by analyzing retrieved business descriptions within our RAG framework, comparing operational language patterns and business model terminology

Context-Aware RAG Implementation

Our methodology leverages context awareness ('"context priming") through a sophisticated prompt-grafting approach that conditions the LLM's responses on specific analytical frameworks. This refers to how our input sequence (the "context") influences the model's conditional probability distribution over the next token. [3]

Our RAG Implementation:



This pipeline ensures that peer set generation is not only automated but also contextually aligned with the nuances of each company's operational footprint. By combining structured metadata, nuanced narrative extraction from filings, and iterative validation through both algorithmic and LLM systems, the approach produces highly relevant and reliable peer groupings while being able to determine the private peers when applicable. This context-aware framework minimizes noise, reduces mismatches, and maximizes analytical precision—enabling more insightful comparisons and robust decision-making across financial, strategic, and competitive analyses.

To showcase how our process creates more insightful comparisons, below is a peer set for Taiwan Semiconductor Manufacturing Company:

3 See Appendix for mathematical basis of our process

Exhibit 1: TSMC Peer Company's Ranked Source: Equity Comparables Model

Target Company	Rank	Peer Company	Public/Private
TSMC	1	GlobalFoundries Inc	Public
	2	United Microelectronics Corp	Public
	3	Semiconductor Manufacturing International Corp	Public
	4	Samsung Electronics Co Ltd	Public
	5	Tower Semiconductor Ltd	Public
	6	Vanguard International Semiconductor Corp	Public
	7	Powerchip Semiconductor Manufacturing Corp	Public
	8	X-FAB Silicon Foundries SE	Public
	9	DB HiTek Co Ltd	Public
	10	Hua Hong Semiconductor Limited	Public
	11	Shanghai Huali Microelectronics Corp	Public
	12	MagnaChip Semiconductor Corp	Public

Source: Morningstar Direct. Data as of April 23, 2026.

Notice that companies like Advanced Micro Devices are not part of the peer set, contrary to their GICS Classification being the same across all levels of granularity.

- TSMC is a pure-play semiconductor foundry; its main business is manufacturing chips for other companies based on their designs.
- AMD is a fabless semiconductor company; it focuses on designing CPUs, GPUs, and other chips, and then outsources the manufacturing, often to TSMC.

While both operate in the semiconductor industry, their roles in the value chain are different: TSMC makes money from wafer fabrication services, while AMD makes money from selling finished processors and graphics chips to consumers and businesses. The ways they generate revenue are fundamentally different, hence our process that looks only for pure play competitors has selected companies that are semiconductor foundries like TSMC.

Multistage Validation Framework

To ensure high-quality peer selections, the methodology implements a comprehensive multistage validation system with parallel processing and expert evaluation.

Response Generation Phase

The system uses parallel generation to produce multiple independent peer analyses for each target company. Multiple response variations are generated simultaneously to reduce processing time while maintaining independence between analyses. Each response follows a standardized JSON format.

Each generated response undergoes multiple validation stages:

- Revenue threshold validation: The system verifies peers having >30% revenue from the target's business model.

- Company Name Validation: The system verifies that peer company names represent actual companies rather than subsidiaries, brands, or product lines.
- A specialized LLM judging panel evaluates each response across multiple criteria.

The system selects the highest-quality response based on:

- Average judge scores across all evaluation criteria
- Consistency of peer selections across multiple response variations
- Adherence to pure-play business model requirements

Current Peer Analysis Limitations

Existing peer company identification methods suffer from several significant limitations that this methodology addresses.

Hierarchical Classification Rigidity

Traditional systems like GICS impose strict hierarchical boundaries that create arbitrary distinctions. A company classified in one sector cannot be easily compared to operationally similar companies in different sectors, even when their business models are substantially similar.

Static Nature

Industry classifications are updated infrequently and struggle to adapt to rapidly evolving business models, particularly in technology and emerging industries where companies may operate across traditional boundaries.

Limited Granularity

Broad industry categories often group companies with fundamentally different business models, revenue sources, and operational characteristics. For example, within "energy," companies focused on renewable energy, oil exploration, and energy infrastructure may have little operational similarity.

Subjective Interpretation

Manual peer selection introduces analyst bias and inconsistency, making it challenging to maintain objective, repeatable criteria across large company universes.

Conclusion

This LLM-based peer company analysis methodology represents a significant advancement over traditional classification-based approaches. By leveraging LLMs to analyze business documents directly, the system can identify operationally similar companies that rigid hierarchical classifications might miss while maintaining strict quality standards through multistage validation.

The system successfully processes thousands of companies while maintaining focus on pure-play peer identification, providing investors and analysts with more accurate and nuanced comparable company analysis than traditional approaches. As business models continue to evolve and cross traditional industry boundaries, this methodology provides the flexibility and accuracy needed for effective peer company analysis in modern markets.

Appendix

Mathematical Foundation for Context Priming

Without RAG, the model generates peer analysis using only its pre-trained parameters:

$$P(peer\_list|\text{Find peers for Apple Inc.}) = softmax(W\_out \cdot h\_final)$$

where $h_f inal$ comes from the model's internal representations of "Apple" and "peers" from training.

Our RAG system fundamentally changes this by injecting retrieved context into the input sequence:

$$Input\ sequence: x = [s^1, s^2, \ldots, s_k, r^1, r^2, \ldots, r_m, q^1, q^2, \ldots, q_n]$$

where:

- s_i= system prompt tokens ("You are a Corporate Strategy Analyst...")
- r_i = retrieved business filings tokens (from vector database similarity search)
- q_i = query tokens ("Find peers for Apple Inc.")

- When generating each token of the peer analysis, the self-attention mechanism computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The key insight is that K and V now include representations of retrieved business descriptions. When generating a peer company name, the attention weights are

$$\alpha_{ij} = \frac{\exp\left(\frac{q_i \cdot k_j}{\sqrt{d_k}}\right)}{\sum_l \exp\left(\frac{q_i \cdot k_l}{\sqrt{d_k}}\right)}$$

where k_j includes keys from retrieved company filing documents. This means the model is attending to retrieved business filings when deciding what peer company to generate next.

The retrieved summaries that mention similar business models (consumer electronics, energy delivery, commercial banking) increase the attention weights α_i for relevant context, making true peers more likely to be generated if the retrieved summaries contain similar operational language.

In our methodology, this enables us to fundamentally shift the model's probability distribution toward generating responses that align with provided specific analytical frameworks. The context about pure-play requirements, revenue thresholds, and operational similarity becomes embedded in the model's attention patterns, influencing how it processes and relates information about different companies.

About Morningstar Quantitative Research

Morningstar Quantitative Research is dedicated to developing innovative statistical models and data points, including the Morningstar Quantitative Rating, the Quantitative Equity Ratings, and the Global Risk Model.

For More Information
Madison.Sargis@morningstar.com

22 West Washington Street
Chicago, IL 60602 USA