Midlincoln Research

February - 10 2025

Analyst: Ovanes Oganisian

Quants

Ranking as Selection: From N-grams and Tokens to Equity Universes

in quantitative equity research, we rank stocks. In large language models, we rank tokens.

This is the third note in a series on quantamental stock rankings. The first note (Ranking Before Prediction) and the second one was (Why Learning Factor Weights Is an Ill-Posed Inverse Problem)

Why ranking keeps reappearing

In quantitative equity research, we rank stocks.
In large language models, we rank tokens.

At first glance, these activities seem unrelated. One deals with financial markets, the other with text. One operates on quarterly or monthly horizons, the other on milliseconds. And yet, once you strip away domain-specific language, the core operation in both systems is the same:

> ***Given a large discrete universe and incomplete information, construct an ordering and act on the top of it.***

Prediction is often presented as the central task. In practice, selection is.

In equities, we rarely need a precise return forecast for every stock. What matters is which stocks end up near the top of the list. In language models, the system does not need to know the “correct” next word. It produces a ranked distribution and selects from it.

This blog explores why ranking-based systems arise naturally across domains, why early language models looked surprisingly similar to factor models, and what modern language models changed—not in the objective, but in *where structure lives*.

Why early language models started with n-grams

Before neural networks dominated language modeling, most systems relied on **n-grams**: sequences of characters or words of fixed length.

A trigram model, for example, estimates probabilities of the form:

> "Given the last two characters, what is the next one?"

From a modern perspective, this approach seems crude. But historically, it was the most natural starting point:

- The alphabet was finite.
- The units were discrete.
- Counting frequencies was tractable.
- The model's behavior was interpretable.

In this sense, n-grams resemble **stock tickers** more than words. They are opaque identifiers, not semantic objects. Their usefulness comes entirely from how often they appear and how they interact with others.

Seen this way, early language models and early factor models share a mindset:

> Define a finite universe of symbols, assign them scores, rank them, and select.

This similarity is structural, not accidental.

Tokens are not words (and not trigrams either)

Modern language models no longer use fixed n-grams. Instead, they rely on **tokens** learned from data.

A token is not a word, and it is not a fixed-length character chunk. It is a **learned text fragment** chosen to balance two competing goals:

- Keep the vocabulary manageable.
- Keep sequences short enough to model long contexts.

Some tokens correspond to whole words. Others are prefixes, suffixes, or frequent word fragments. Many include whitespace or punctuation. Their defining property is not linguistic purity but **statistical efficiency**.

The cleanest way to think about tokenization is as a form of **compression**:

- Frequent substrings get their own symbols.
- Rare substrings are broken into smaller pieces.

- The tokenizer builds a dictionary that minimizes overall description length under a size constraint.

Tokens are therefore best understood as **handles**, not meanings.

This matters because it prevents a common misunderstanding: tokens are not the “atoms” of language. They are just a convenient alphabet over which ranking and selection can occur.

Where the analogy almost breaks—and why it still holds

At this point, an important difference becomes apparent, but it is not the one most analogies focus on.

In language models, the candidates being ranked—tokens—are also the building blocks of the output. A sentence is literally constructed by selecting tokens one after another. Ranking and construction are the same operation.

In equity ranking systems, the candidates being ranked are stocks. The output is not constructed from stocks in the same sense; it is an ordering or allocation over a universe of already-existing objects. One cannot “build” a stock the way one builds a sentence.

Taken at face value, this seems to break the analogy. But the difference is subtler.

When we use simple financial descriptors—fundamental ratios, growth metrics, balance-sheet statistics, price-based signals—the immediate output of the model is a ranked set of tickers. In this sense, the ranking itself is already a **constructed object**: a low-dimensional representation of a much richer financial context. Much like an embedding in a language model, the ranking compresses information rather than expressing it directly.

This naturally raises the question: can we introduce financial “tokens” analogous to those used in modern language models?

If we attempt this, the candidate tokens are not stocks themselves, but placeholders for informational structure: earnings outcomes, policy decisions, macroeconomic shifts, political events, narrative changes, and other sources of uncertainty. These tokens do not form outputs. Instead, they participate in the **construction of context**.

Here the analogy changes form.

In language models, tokens directly compose both context and output. In financial systems, tokens—if they exist at all—can only compose **context**, not portfolios. The constructed context then induces a ranking over stocks.

This distinction matters. Financial context can be compressed and represented, but the rules of construction differ. Tokens do not assemble into assets; they assemble into a state against which assets are evaluated.

The correct distinction, therefore, is not between “construction” and “selection,” but between two types of construction:

- In language, construction operates on the same alphabet as the output.
- In finance, construction operates on a latent context, and the output is a ranking over a separate universe.

Once this is made explicit, the analogy becomes precise rather than sloppy. Both systems rely on ranking. Both systems compress rich information into a form that makes ranking possible. They simply differ in *where* construction stops and selection begins.

What “context” actually means

Another common source of confusion is the role of “events” or “signals.”

When we say that earnings, policy decisions, or narratives matter, we do **not** mean that these are fixed input variables in a clean feature matrix. They are placeholders for **uncertain informational perturbations**.

In language models, the system does not know in advance which words will matter. It processes a stream of symbols, forms a latent context, and ranks possible continuations.

In markets, we face the same structure:

- Some information arrives.
- Some does not.
- Some matters only in certain regimes.
- Meaning is revealed only after interaction.

The important object is not the event itself, but the **latent context** it induces.

Language models make this explicit: context is a continuously updated internal representation. In finance, context is often handled implicitly—through heuristics, regime switches, or discretionary judgment—but the role is the same.

Context does not predict outcomes.
Context **tilts rankings**.

What modern language models actually changed

The leap from n-grams to modern language models was not about bigger alphabets or better optimization. It was about **where structure lives**.

Early models tried to encode structure in the symbols themselves:

- Fixed-length n-grams.
- Hard-coded backoff rules.
- Local dependencies.

Modern models moved structure into **representations**:

- Tokens are cheap identifiers.
- Meaning is context-dependent.
- The same token can behave very differently in different states.

This shift has a direct analogue in equity modeling.

Static factor weights assume that the meaning of “value,” “growth,” or “quality” is constant. In reality, these concepts are regime-dependent. Their influence changes with liquidity, rates, sentiment, and cross-sectional structure.

Modern language models succeed not because they rank better, but because they rank **in a better space**.

Why this matters for equity ranking systems

Seen through this lens, many practical observations fall into place:

- Why factor weights feel unstable.
- Why many weight vectors produce similar rankings.
- Why middle-ranked stocks are interchangeable.
- Why top-k stability matters more than exact ordering.

Linear factor models are not wrong; they are **low-context** approximations. They assume a fixed geometry and ask the ranking to do all the work.

Language models invert this: they enrich the geometry so that ranking becomes easier.

The lesson for finance is not to copy neural architectures wholesale, but to recognize that:

> **Ranking quality depends more on representation than on optimization.**

Ranking as the invariant object

Across domains, the same object keeps appearing:

- A finite universe.
- Incomplete information.
- A need to act on a small subset.
- Evaluation that is delayed or noisy.

In all these cases, ranking emerges as the natural operation.

Language models make this explicit and mechanistic. Equity systems arrived there through practice.

Once ranking is recognized as the invariant object, debates about prediction versus classification, or AI versus traditional models, become secondary. The real questions shift to:

- Which rankings are even possible?
- How stable are they?
- How sensitive are they to context?

Those questions are not philosophical. They are geometric.

And that is where the next blog turns.

Closing

This note did not argue that language models and equity ranking systems are the same. They are not.

It argued something narrower and more defensible:

> Both systems solve a selection problem under uncertainty by inducing an ordering over a discrete universe.

Language models taught us that the hardest part of such systems is not ranking itself, but designing representations in which ranking becomes meaningful.

In finance, we have been solving the same problem all along—often without naming it.

The next step is to ask a sharper question:

> **What rankings can linear scoring rules actually produce—and which are impossible no matter how we tune them?**

That is not a modeling choice.
It is a mathematical constraint.

The information contained in this document is neither an offer to sell nor a solicitation of an offer to purchase interests in any referenced investment nor does it represent a research report. Securities may not be offered or sold in the United States absent registration with the US Securities and Exchange Commission or an exemption from registration under the US Securities Act of 1933, as amended. This document is only directed at professional investors who have experience of investing in emerging markets and the referenced investments are unlikely to be suitable for most private individuals. The referenced investments are speculative and include a high level of risk, and investors may not receive back the original amount of money that they invested. The value of investments can fall as well as rise, and you may get back less than what you originally invested. Where an investment is made in overseas currencies, changes in currency exchange rates may affect the value of your investment. Investments in emerging markets can be more volatile than in other more developed markets. Past performance is no guarantee of future performance, and the value of investments can go down as well as up. Please consult your financial and tax advisers if you are considering investing in any of the referenced investments. This document may contain certain forward-looking statements with respect to MidLincoln Research's strategies or expectations. Forward-looking statements speak only as of the date they are made, and MidLincoln Research assumes no duty to, and does not undertake to, update such forward-looking statements. This document may not be reproduced, distributed, transmitted, displayed, published or broadcast by any recipient for any purpose without the prior consent of MidLincoln Research. This document has been issued by MidLincoln Research.