

Asset and Investor Embeddings*

Xavier Gabaix[†] Ralph S.J. Koijen[‡] Robert J. Richmond[§] Motohiro Yogo[¶]

July 16, 2026

Abstract

Traditional representations of firms use accounting and financial market data, but investors use richer information sets. Theoretically, we show that portfolio holdings contain all relevant information for the cross-section of asset prices, recoverable under empirically realistic conditions. Building on recent advances in machine learning and artificial intelligence, we develop asset embeddings as learned firm representations from portfolio holdings. We evaluate different methods of estimating asset embeddings on three new benchmarks that can be evaluated with a single cross section of data. We also develop investor embeddings as learned representations of investors and their strategies with applications in performance measurement and detection of crowded trades. We provide economic interpretations of asset (investor) embeddings by applying large language models to firm- (investor-)level text data.

JEL: C53, G12, G23

Keywords: Artificial intelligence, Asset pricing, Benchmarks, Machine learning, Recommender systems, Transformer models

*For comments and discussions, we thank Tania Babina, Bernard Bergeron, Svetlana Bryzgalova, Leland Bybee, John Campbell, Kent Daniel, Tarek Hassan, Bryan Kelly, Serhiy Kozak, Martin Lettau, Tengyuan Liang, Stefan Nagel, Sangmin Oh, Markus Pelger, Fernando Perez-Cruz, Jerome Pesenti, Tarun Ramadorai, Olivier Scaillet, Hyun Shin, Yinan Su, and participants at seminars and conferences. We thank Manav Chaudhary, San Singh, and Huanyu Zhang for excellent research assistance. For financial support, Gabaix thanks the Ferrante Fund, and Koijen thanks the Center for Research in Security Prices at the University of Chicago and the Fama Research Fund at the University of Chicago Booth School of Business. This paper previously circulated with the title “Asset Embeddings.”

[†]Harvard University, NBER, and CEPR (xgabaix@fas.harvard.edu)

[‡]University of Chicago Booth School of Business, NBER, and CEPR (ralph.koijen@chicagobooth.edu)

[§]New York University Stern School of Business and NBER (rrichmon@stern.nyu.edu)

[¶]Princeton University and NBER (myogo@princeton.edu)

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have revolutionized our understanding of complex unstructured data such as text and images by representing these data with high-dimensional latent vectors called embeddings. The embeddings are estimated by self-supervised training methods that do not require labeled data. In finance and economics, representations of firms, assets, and investors play a central role in many fields, including asset pricing, corporate finance, macroeconomics, and industrial organization. However, instead of learned representations, economists have traditionally relied on firm characteristics, derived from accounting and financial market data, as well as investor attributes such as type, style, and size, to explain complex economic outcomes. We introduce asset and investor embeddings, learned from portfolio holdings data, to represent assets and investors.

There are numerous data sources that we can use to learn representations, including accounting, financial market, and text data (e.g., 10-K filings and earnings calls). Our first contribution is to show in Section 2 that, under generic conditions, portfolio holdings contain all relevant information about firms for asset prices. Because asset prices are determined by an asset demand system and market clearing, portfolio holdings data contain all information that determines asset prices under empirically-relevant conditions. The specification of the asset demand system determines how we estimate embeddings from portfolio holdings data, which motivates us to explore different models. For example, a log-linear asset demand system implies a linear factor model to estimate embeddings. Based on our theoretical result, we primarily focus on portfolio holdings data as a natural starting point. Empirically, we find that portfolio holdings data are not spanned by traditional characteristics and simple text representations (and vice versa). Of course, this finding does not rule out other data sources to represent assets or investors, particularly in small samples. Learning richer representations using observed characteristics, text, and portfolio holdings is an exciting direction for future research.

Our second contribution is to use various models to estimate asset and investor embeddings with portfolio holdings data, building on the self-supervised learning paradigm in the NLP literature. Our central insight is that portfolio holdings data provide a natural analogue to text, allowing methods from the NLP literature to be adapted to finance and economics. Just as documents structure words in ways that allow us to estimate word embeddings, investors structure portfolios in ways that allow us to estimate asset embeddings. This logic allows us to build on models traditionally used for text and estimate them instead using portfolio holdings. We consider three classes of models. The first is recommender systems, which include principal component analysis (PCA) models. The second is shallow neural network

models such as Word2Vec. The third is transformer models used in large language models (LLMs) such as the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), which is our main focus, and the Generative Pre-Trained Transformer (GPT) (Radford et al., 2018).

To adapt Word2Vec and the BERT models to our application, we need to convert numerical values of portfolio holdings to discrete words that these models use as inputs. We order the assets in an investor’s portfolio in a decreasing order of portfolio weights and simply represent the name of each asset as a token. This procedure exploits the fact that investors assign similar portfolio weights to assets with similar asset embeddings, according to the portfolio theory underlying the asset pricing model in Section 2. We then train the BERT model to predict a masked asset (e.g., the second largest position), given the other assets in an investor’s portfolio. This training procedure is analogous to the training of Word2Vec and BERT models by self-supervised masked language modeling. For example, the BERT model has to predict the missing word in a sentence like “The Fed decided to ____ the target rate to fight inflation.”

Our third contribution is to introduce three benchmarks to evaluate the various models of asset embeddings. First, we predict relative valuations in the cross section of firms. Second, we explain the comovement of stock returns. Third, we predict masked assets in the managed portfolios of mutual funds, exchange-traded funds (ETFs), and hedge funds. This benchmark tests a core prediction of the portfolio theory that investors assign similar portfolio weights to assets with similar asset embeddings. Our benchmarks satisfy three important criteria. First, the asset pricing model in Section 2 predicts that portfolio holdings data are informative for these benchmarks. Second, these benchmarks are at the heart of asset pricing and portfolio theory, for which we have existing models (e.g., the characteristics of the Fama-French five-factor model) that establish a high hurdle. Third, these benchmarks require only a single cross section of data (e.g., just one quarter). This criterion facilitates a true out-of-sample test to measure progress in real time. For example, we could imagine an annual competition to evaluate new models on these benchmarks, just like the ImageNet Large Scale Visual Recognition Challenge. Traditional benchmarks in asset pricing tend to focus on average returns, which can only be measured with precision over long periods of time. While those benchmarks are important, the cross-sectional benchmarks that we introduce are equally relevant as a way to test asset pricing and portfolio formation models, and allow for faster iteration.

Our fourth contribution is to estimate and to evaluate the various models of asset embeddings. We use US stock holdings data for mutual funds, ETFs, closed-end funds, variable annuity funds, and hedge funds over a quarterly sample period from 2005.Q1 to 2022.Q4.

Based on the three benchmarks, we compare the performance of the asset embeddings, firm characteristics, and text-based embeddings from Cohere and OpenAI, which are leading AI companies. Three main insights emerge from this analysis.

First, all models of asset embeddings significantly outperform firm characteristics on the relative valuation and managed portfolio benchmarks. For example, a base model of four-dimensional asset embeddings explains over 50% of the variation in relative valuations, whereas a set of four leading firm characteristics explains only 15%. The performance of the asset embeddings easily scales in their dimension. When we train a large model of 128-dimensional asset embeddings by ridge regression with cross validation, we explain over 70% of the variation in relative valuations. A base model of six-dimensional asset embeddings performs well on the return comovement benchmark, but it falls slightly short of a high hurdle of a set of characteristics in the Fama-French five-factor model and momentum. As before, the performance improves when we use higher-dimensional embeddings and this model surpasses the five-factor model.

Second, we find important performance gains from using different models for different tasks. The recommender systems perform well on the relative valuation and return comovement benchmarks. However, Word2Vec and the BERT model significantly outperform the recommender systems on the managed portfolio benchmark. This result makes sense since we train Word2Vec and the BERT model to predict masked assets in managed portfolios. The BERT model performs particularly well on the managed portfolio benchmark, which highlights the importance of contextualized embeddings for predicting institutional portfolio decisions.

Third, the text-based embeddings from Cohere and OpenAI perform poorly on our benchmarks. We show that this is partly caused by the fact that text embeddings mix semantic and economic similarity. That said, one can plausibly improve the performance of text-based embeddings by changing the source of the text data or fine-tuning embeddings models. More broadly, future research could use a combination of portfolio holdings data, a large set of firm characteristics, and a collection of text data to estimate asset embeddings that most accurately represent firms. This approach mirrors recent developments in the computer science literature that focus on world models (Balestriero and LeCun, 2025). We show, both theoretically and empirically, that portfolio holdings data are essential for any attempt to represent firms and investors, which is not the current practice at even the leading AI labs.

We use the same models to estimate investor embeddings, which are vector representations of investors and their strategies. The investor embeddings provide a richer representation of investors than standard investor characteristics such as institutional type (e.g., mutual fund, hedge fund, or insurance company), size, and measures of activeness (e.g., turnover

and active share). We use a BERT model to estimate the investor embeddings and show that it performs well on a masked investor benchmark. We discuss several applications of investor embeddings, including investor classification, performance evaluation, and detecting crowded trades.

Our fifth contribution is to develop a framework that leverages recent advances in LLMs to interpret asset and investor embeddings (Section 8). For asset embeddings, we first feed the earnings calls of each firm to an LLM and ask for themes that may appeal to investors, such as a firm’s growth prospects, strong brand value or corporate policies that may be attractive to certain groups of investors. Then, for any set of firms that have similar holdings-based asset embeddings, we prompt the LLM to find the common themes based on these text-based summaries. This provides an economic narrative why investors hold a given set of stocks. We apply the same methodology to investor embeddings. Instead of using earnings calls, we use text data from (mostly) prospectuses in which investors describe their strategies and investment philosophies. By identifying similarities in the strategy descriptions for investors that are close in investor embeddings space (based on the holding data), we provide an economic interpretation of investor embeddings as well. For the textual analysis, we use the Gemini 2.5 Pro model from Google, which is a state-of-the-art reasoning model as of June 2025. To illustrate the narratives beyond the examples provided in the paper, readers can access our results based on the latest regulatory filings on a companion website to this paper (<https://www.mktgen.ai>). Here, we also use the latest language models from the Gemini series to interpret the embeddings.

Finally, Section 9 proposes several applications of asset and investor embeddings for future research. First, we could use the asset embeddings to construct replicating portfolios without historical data on stock returns. For example, we could start with a strategy that is long United Airlines and short Zoom, which are exemplar stocks for exposure to COVID-19 risks. To construct a diversified COVID-19 factor, we could use the asset embeddings to find stocks that are most similar to United and Zoom. Second, financial institutions, regulators, and central banks could use asset and investor embeddings for risk management and stress testing. They could generate counterfactual changes in asset and investor embeddings, which imply counterfactual changes in asset prices in an asset demand system (Kojien and Yogo, 2019). Thus, we propose a novel methodology to generate stress scenarios, which could have broad application in risk management.

In summary, we provide five main contributions. First, we provide a theoretical foundation for using portfolio holdings to recover asset and investor embeddings. Second, we recover the embeddings using various embedding models, providing a new connection between the NLP and finance literature. Third, we introduce three new benchmarks to evaluate em-

beddings using a single cross section of data. Fourth, we evaluate the different embedding methods on these benchmarks and compare the results to observed characteristics and text embeddings as a point of reference. Fifth, we show how to interpret asset and investor embeddings using text data from earnings calls and investor prospectuses, which provides an important degree of interpretability of our results.

Related Literature A large literature uses a large number of firm characteristics to model the stochastic discount factor or to predict stock returns (Kelly et al., 2019; DeMiguel et al., 2020; Feng et al., 2020; Freyberger et al., 2020; Kozak et al., 2020; Lettau and Pelger, 2020; Cong et al., 2022). We refer to Kelly and Xiu (2023) for a review of this literature. However, firm characteristics are only a subset of the investors’ information set (Hansen and Richard, 1987). Since portfolio holdings contain all relevant information about firms, we can in theory recover the full information set by estimating asset embeddings from portfolio holdings data, while no such guarantees for other data sources like accounting and text data.

Another literature uses text data to represent firms and to define firm similarity, including 10-K filings (Hoberg and Phillips, 2016; Chen and Sarkar, 2020), earnings calls (Hassan et al., 2019), and news articles (Binsbergen et al., 2023; Bybee et al., 2023; Sarkar, 2025). Other sources of economic linkages between firms include customer-supplier relations (Cohen and Frazzini, 2008), production networks (Lee et al., 2024), and shared analyst coverage (Ali and Hirshleifer, 2020). We ultimately view the characteristics and text data as complementary to the portfolio holdings data. Moreover, we can estimate the asset embeddings with the portfolio holdings data and give them an economic interpretation with the text data, as we demonstrate in Section 8.

A recent literature estimates factor models on portfolio holdings data for various purposes (Madhavan et al., 2021; Gabaix and Koijen, 2023; Betermier et al., 2025; Balasubramaniam et al., 2023; Clayton and Coppola, 2025). We extend this literature by the five contributions listed above.

Most of our analysis builds on recent advances in AI that offer valuable tools to process unstructured data such as text, audio, and images. The most relevant tools for this paper are NLP models (Bengio et al., 2003; Mikolov et al., 2013a,b; Pennington et al., 2014) and transformer models (Vaswani et al., 2017; Devlin et al., 2019). In an application to the stock market, Dolphin et al. (2022) estimate a neural network model to capture the covariance structure of daily stock returns. This paper is part of a growing literature that applies ML to economics and finance (Hassan et al., 2019; Gu et al., 2020; Nagel, 2021; Bybee et al., 2023; Chen et al., 2024; Hommel et al., 2023). A branch of this literature studies missing firm characteristics in asset pricing applications (Bryzgalova et al., 2025; Freyberger et al.,

2022; Beckmeyer and Wiedemann, 2023; Chen and McCoy, 2024).

This paper contributes to the literature on demand system asset pricing (Kojien and Yogo, 2019). This literature develops theoretical and empirical frameworks to study asset prices, firm characteristics, and macro variables together with portfolio holdings. By substituting the equilibrium asset prices in asset demand, we derive reduced-form demand. We can estimate asset and investor embeddings from reduced-form demand without instruments or comprehensive holdings data that satisfy market clearing. Thus, we develop a recipe for prediction exercises using an asset demand system, which are possible under weaker identifying assumptions than those required for counterfactual exercises.

Notation We index the investors by $i = 1, \dots, I$, the assets by $a = 1, \dots, A$, and time by t . Time is at a quarterly frequency in our applications, which are based on quarterly portfolio holdings data. When we use monthly stock returns for the return comovement benchmark, we index time at a monthly frequency by m . Since our application is the stock market, we use the words “asset”, “stock”, and “firm” interchangeably throughout the paper. For concepts that involve only cross-sectional data, we omit the time subscript t for simplicity.

We use calligraphic letters such as $\mathcal{A}$ and $\mathcal{B}$ to denote sets with the corresponding number of elements $|\mathcal{A}|$ and $|\mathcal{B}|$. We denote the set of assets in investor i ’s portfolio as $\mathcal{A}_i$ with the corresponding number of assets $|\mathcal{A}_i|$. We order the assets in investor i ’s portfolio in a decreasing order of portfolio weights, where $a_i(k)$ is the name of the k -th largest position. Thus, investor i ’s assets are $a_i(1), a_i(2), \dots, a_i(|\mathcal{A}_i|)$. We denote the set of assets in investor i ’s portfolio, excluding asset $a_i(k)$, as $\mathcal{A}_i(k) = \mathcal{A}_i \setminus \{a_i(k)\}$. We denote the set of investors that hold asset a as $\mathcal{I}_a$ with the corresponding number of investors $|\mathcal{I}_a|$. We order the investors that hold asset a in a decreasing order of ownership shares, where $i_a(k)$ is the name of the k -th largest investor. Thus, asset a ’s investors are $i_a(1), i_a(2), \dots, i_a(|\mathcal{I}_a|)$.

For each asset, we normalize its shares outstanding to one. We denote the price of asset a at time t as P_{at} , which is also its market capitalization (with the shares outstanding normalized to one). We denote the number shares of asset a that investor i holds at time t as Q_{iat} . Thus, the dollar holding is $H_{iat} = P_{at}Q_{iat}$. We use a lowercase letter to denote the logarithm of the corresponding uppercase variable. For example, $h_{iat} = \ln(H_{iat})$. We use bold letters to denote column vectors and matrices. For example, we denote a vector of asset embeddings as $\mathbf{x}_a = [x_{ak}]_{k=1}^K \in \mathbb{R}^K$ and the corresponding matrix of asset embeddings as $\mathbf{x} = [\mathbf{x}'_a]_{a=1}^A \in \mathbb{R}^{A \times K}$. We denote the Frobenius norm of the matrix as $\|\mathbf{x}\| = \left(\sum_{a=1}^A \sum_{k=1}^K x_{akt}^2 \right)^{\frac{1}{2}}$.

2. Portfolio Holdings as Embeddings Data

We show in Section 2.1 that, under very general conditions, holdings data are a natural source of data to estimate asset and investor embeddings. We specialize the model in Section 2.2 using a log-linear asset demand system to provide an illustrative example. This model microfound the use of recommender systems to estimate embeddings. We only present the main equations and leave all proofs for Appendix A.

2.1. The general logic

We write an investor's vector of asset demand $\mathbf{Q}_i = (Q_{ia})_{a=1\dots A}$ as dependent on the vector of prices $\mathbf{P}$, and the matrix of other asset characteristics (later, embeddings) $\mathbf{x}$:

$$\mathbf{Q}_i = \mathbf{Q}_i(\mathbf{P}, \mathbf{x}), \quad (1)$$

where we omit time subscripts for simplicity. This general formulation allows the demand for stock a to depend on its own price and characteristics, but also the prices and characteristics of all other stocks. In a dynamic model, lagged prices and characteristics may influence demand as well.

Market clearing implies $\sum_i \mathbf{Q}_i = \mathbf{S}$, where $\mathbf{S}$ denotes the vector of asset supplies. We allow for the possibility that supply is elastic, $\mathbf{S}(\mathbf{P}, \mathbf{x})$. We assume that the demand system has a unique solution, which implies

$$\mathbf{P} = \mathbf{b}(\mathbf{x}). \quad (2)$$

By substituting prices into (1), it holds

$$\mathbf{Q}_i = \mathbf{Q}_i(\mathbf{b}(\mathbf{x}), \mathbf{x}) =: \tilde{\mathbf{Q}}_i(\mathbf{x}). \quad (3)$$

We refer to $(\mathbf{Q}_i(\mathbf{P}, \mathbf{x}))_{i=1}^I$ as the structural demand system and $(\tilde{\mathbf{Q}}_i(\mathbf{x}))_{i=1}^I$ as the reduced-form demand system.

Our simple, yet general, insight is that all information that is relevant for asset prices, see (2), is generically embedded in the reduced-form demand system, see (3). Different demand systems imply different mappings from embeddings to holdings and prices. For this reason, we explore different methods to recover asset and investor embeddings. In the next section, we provide an illustrative example for a log-linear demand system.

2.2. An illustrative example: A log-linear asset demand system

We model investor i 's log dollar holding of asset a as

$$h_{ia} = \kappa_i + (1 - \zeta_i) p_a + \nu_{ia}, \quad (4)$$

where κ_i is a constant that scales with the investor's wealth and ζ_i is the price elasticity of demand for individual assets.¹ The demand shifter has a factor structure:

$$\nu_{ia} = \mathbf{\Lambda}'_i \mathbf{x}_a + \xi_{ia}. \quad (5)$$

The K -dimensional vector $\mathbf{x}_a$ is an asset embedding that contains all relevant information about asset a . This information includes firm characteristics, text data, and other sources of information that may not be readily available. The K -dimensional vector $\mathbf{\Lambda}_i$ is the semi-elasticity of the portfolio holding to the asset embedding, conditional on the asset price. Investors respond heterogeneously to different elements of the asset embedding. For example, some investors use text data, but other investors do not. The scalar ξ_{ia} represents the idiosyncratic component of investor i 's demand for asset a .

Koijen and Yogo (2019) derive equations (4) and (5) in a traditional model of portfolio choice when the returns have a factor structure, the factor loadings depend on firm characteristics, and investors have heterogeneous beliefs and agree to disagree. Koijen et al. (2024) extend this microfoundation to models of portfolio choice with hedging demand and non-pecuniary preferences for firm characteristics (e.g., ESG scores). Under this microfoundation, the asset embeddings capture differences in expected profitability or risk exposure across assets. Equation (4) is linear in log asset price, and equation (5) is linear in the asset embedding. These equations can be nonlinear in more general models of portfolio choice with constraints.

In Appendix A, we solve for equilibrium asset prices by market clearing. We then substitute the equilibrium asset prices in asset demand (4) to derive reduced-form demand:

$$h_{ia} = \boldsymbol{\lambda}'_i \mathbf{x}_a + \delta_i + \delta_a + \epsilon_{ia}. \quad (6)$$

The K -dimensional vector $\boldsymbol{\lambda}_i$ is an investor embedding, which is a linear transformation of $\mathbf{\Lambda}_i$ and ζ_i . The investor embedding $\boldsymbol{\lambda}_i$ in reduced-form demand (6) represents the investor's preferences for asset embeddings, relative to the other investors' preferences. In contrast,

¹In the language of Gabaix and Koijen (2023), ζ_i is the micro elasticity (i.e., substitution across individual stocks) rather than the macro elasticity (i.e., substitution between the aggregate stock market and the aggregate bond market).

the demand elasticities Λ_i in asset demand (5) represent the absolute preferences for asset embeddings, which require structural estimation by instrumental variables (Koijen and Yogo, 2019). The investor fixed effect δ_i controls for differences in wealth across investors. The asset fixed effect δ_a controls for differences in market capitalization across assets. Given these fixed effects, equation (6) models the active deviations of the portfolio weights from the market weights.

We can estimate the asset embeddings in reduced-form demand (6) by PCA. A necessary condition for identification is heterogeneity in λ_i (through either Λ_i or ζ_i) across investors, which is the mathematically generic, and indeed empirically relevant case with heterogeneity in the investment strategies.²

Investors with very different investment strategies (e.g., that disagree more) are especially informative for identifying asset embeddings. Thus, an econometrician can asymptotically recover the same information set as investors. In finite samples, firm characteristics and text data can augment the explanatory power of asset embeddings in actual applications due to estimation error in estimating the asset embeddings.

In Appendix A, we make an additional assumption that the asset embeddings are stable from time $t - 1$ to t . Then the same asset pricing model implies that we can use returns, volume, or portfolio rebalancing (i.e., change in shares held) as alternative sources of embeddings data.

3. Applying AI Methods to Economics and Finance

Investors collectively use all relevant information about firms, including firm characteristics, text data, and other sources of information that are not readily available. Building on the insight that the portfolio holdings contain all relevant information, we explore various methods to extract this information from portfolio holdings data. Recent advances in processing text, audio, and images offer valuable AI tools for this purpose. In particular, NLP models represent words with high-dimensional latent vectors called embeddings. Embeddings capture similarities between words, sentences, and documents and can do math with words. In a classic example, the embeddings for “king”, “man”, and “woman” are used to compute “king – man + woman”, which produces an embedding closest to “queen”.

In economics and finance, we commonly use firm characteristics, based on accounting and financial market data, to model the similarity between firms. For example, we use firm characteristics to understand relative valuations, comovement of stock returns, and institutional

²Formally, we recover all the information $(\mathbf{x}_a)_{a=1\dots A}$ if $\mathbf{\Lambda}'\boldsymbol{\lambda}$ has full rank K , where $\boldsymbol{\lambda} \in \mathbb{R}^{I \times K}$ is the matrix of reduced form holdings. This assumption would not hold with a representative investor. Fortunately, generically, investors are heterogeneous.

portfolio decisions. The AI literature suggests a different approach. Since the portfolio holdings contain all relevant information about firms, we can extract this information directly from portfolio holdings data using modern AI methods. Our central insight is simple yet powerful. Just as documents structure words in NLP, songs structure notes in audio signal processing, and images structure pixels in computer vision, portfolios structure firms in economics and finance.

3.1. From Words and Sentences to Assets and Investors

We summarize developments in NLP over the last three decades that are relevant for our application to economics and finance. Our high-level overview focuses on the conceptual issues and is not meant to be exhaustive. For textbook treatments of NLP, we refer to Prince (2023) and Jurafsky and Martin (2025).

We start with a simple yet powerful class of NLP models called recommender systems, which include PCA models and latent semantic analysis (Dumais et al., 1988). In a simple recommender system, we apply PCA to the document-word matrix to estimate word embeddings. Similarly, we apply PCA to the investor-asset matrix of portfolio holdings to estimate asset embeddings.

The next important step in the NLP literature was shallow neural network models such as Word2Vec (Mikolov et al., 2013a,b). In Word2Vec, we represent words as embeddings and train a shallow neural network to predict a target word, based on the surrounding words in a context window.³ Thus, Word2Vec is a local nonlinear model in contrast to recommender systems, which are global bilinear models. To apply Word2Vec to portfolio holdings data, we order the assets in an investor’s portfolio in a decreasing order of portfolio weights. According to the portfolio theory underlying equation (6), investors assign similar portfolio weights to assets with similar asset embeddings. We then train the model to predict a target asset (e.g., Apple), based on similarly-sized positions in the context window (e.g., Alphabet, Amazon, and Microsoft).

Word2Vec has two important limitations. First, the order of words in the context window does not matter. Second, the embeddings are context-invariant and have a one-to-one mapping to words. The second limitation is also relevant for our application. For example, Apple produces computer hardware (e.g., M4 chips), provides digital content and services (e.g., iCloud and Apple Music), and innovates in AI (e.g., Apple Intelligence). Thus, we should view Apple as a computer hardware company in a portfolio that includes Dell and Intel but as a digital content provider in a portfolio that includes Netflix and Spotify.

³This method is the continuous bag of words. Alternatively, the continuous skip-gram method uses the target word to predict the surrounding words in a context window.

Finally, we consider transformer models such as BERT and GPT, which address the limitations of Word2Vec (Vaswani et al., 2017). A transformer model estimates a contextualized embedding as a weighted average of the word embeddings in a sentence. In our application, the transformer model estimates a contextualized embedding as a weighted average of the asset embeddings in a portfolio. The transformer model also uses information about the order of the assets in the entire portfolio through position embeddings.

3.2. Recommender Systems

Starting with reduced-form demand (6), we apply PCA to log dollar holdings. We estimate the parameter vector $\theta = (\lambda'_i, \mathbf{x}'_a, \delta_i, \delta_a)'$ through the objective function:

$$\min_{\theta} \sum_{i=1}^I \sum_{a=1}^A (h_{ia} - \lambda'_i \mathbf{x}_a - \delta_i - \delta_a)^2 + \frac{\eta_{\lambda}}{IK} \sum_{i=1}^I \lambda'_i \lambda_i + \frac{\eta_x}{AK} \sum_{a=1}^A \mathbf{x}'_a \mathbf{x}_a. \quad (7)$$

The first term is the mean squared error for log dollar holdings. The second and third terms regularize the investor and asset embeddings.

A recommender system won the Netflix Prize, which was a 2006 competition to predict whether a viewer i would like a movie, given her past ratings h_{ia} on a subset of movies. The embedding $\mathbf{x}_a$ represents the latent characteristics of movie a , and the embedding λ_i represents the viewer’s preferences over the latent characteristics. Consider an alternative approach based on observed characteristics such as genre, length, and the parental guidance rating. Observed characteristics would have difficulty predicting preferences for niche categories like teenage vampire movies that are romantic. A recommender system automatically generates embeddings that represent “teenage”, “vampire”, and “romantic” and delivers an accurate recommendation.

Similarly, suppose that some investors like firms with a large number of GPUs for AI computing, high intangible capital, and generous work-from-home policies. This type of information is not readily available in accounting data and is difficult to systematically collect for all firms. However, a recommender system can recover this information from portfolio holdings data, as long as some investors collect and use this information in portfolio choice. Thus, asset embeddings could have broader economic application to represent firm characteristics that are difficult to measure, such as intangible capital or work-from-home policies.

Variants We consider several variants of the recommender system to learn which dimensions of the portfolio holdings data are most useful for the various benchmarks in Section

4. We take an exploratory approach in the absence of theoretical predictions on which variant performs best on which benchmark. The portfolio holdings data are sparse with many zero positions. In a variant labeled RS-Binary, we use only binary information for positive versus zero positions. In equation (7), we set $h_{ia} = 1$ for positive positions, $h_{ia} = 0$ for zero positions, and $\delta_i = 0$. We estimate this variant by PCA without regularization (i.e., $\eta_\lambda = \eta_x = 0$).

In a variant labeled RS-Ranks, we use only percentile ranks and replace zero positions with zeros. In equation (7), we set h_{ia} to the percentile rank (i.e., ranking of the positive positions divided by $|\mathcal{N}_i|$) such that $h_{ia} = 1$ for the largest positive position, $h_{ia} = \frac{1}{|\mathcal{N}_i|}$ for the smallest positive position, and $h_{ia} = 0$ for the zero positions. We estimate this variant by PCA without regularization.

We next consider two variants that use log dollar holdings but make different economic assumptions about the zero positions. We first estimate the investor fixed effects as the mean of h_{ia} by investor. We then estimate the asset fixed effects as the mean of $h_{ia} - \delta_i$ by asset. We then remove the investor and asset fixed effects as $\hat{h}_{ia} = h_{ia} - \delta_i - \delta_a$. In a variant labeled RS-L-0, we replace the zero positions with zeros (i.e., the average value conditional on investor and asset fixed effects). This variant assumes that the zero positions are uninformative, as if they were determined by investment mandates. In a variant labeled RS-L-Min, we replace the zero positions with the smallest value of $\hat{h}_{ia}$ for investor i . This variant assumes that the zero positions reflect a negative view, as if they were determined by binding short-sale constraints. We estimate both variants by PCA without regularization.

In a final variant labeled RS-L, we use only the intensive margin of positive positions. In equation (7), we use log dollar holdings but sum over only positive positions (i.e., $\sum_{a \in \mathcal{N}_i}$). We estimate the regularization parameters (η_λ, η_x) by ten-fold cross-validation. We refer to Appendix B for further details on the estimation.

We discuss two potential extensions. First, we estimate high-dimensional unsupervised asset embeddings and train them on our benchmarks by ten-fold cross validation. Since we find good performance on all of our benchmarks, we did not pursue supervised estimation in this paper. However, supervised estimation is a straightforward extension that may be a powerful alternative, which produces lower-dimensional asset embeddings that are specialized to the benchmark (Bryzgalova et al., 2023). We could estimate the asset embeddings through a supervised objective function (B1), estimating the regularization parameters by ten-fold cross validation. Second, in Appendix C, we propose to use historical data, in addition to the current quarter, to sharpen the embedding estimates.

3.3. Shallow Neural Network Models

Word2Vec exploits the fact that related words appear closely together in a sentence such as “salt” and “pepper”. In analogy, investors assign similar portfolio weights to assets with similar asset embeddings, according to the portfolio theory underlying equation (6). Therefore, we order the assets in an investor’s portfolio in a decreasing order of portfolio weights.⁴ In the training data, we mask the k -th largest position in each investor’s portfolio, which we denote as $b_i(k)$. We then train Word2Vec to predict the masked asset, based on the surrounding assets in a context window. We denote the set of assets in the context window as $\mathcal{C}_i(k) = \{a_i(l) \mid 0 < |k - l| \leq \chi\}$, where χ determines the size of the context window.

Let $\mathcal{B} = \{b_i(k) = [\text{MASK}]\}$ be the set of all masked assets in the validation data, where the true identity of the masked asset is $b_i(k) = a_i(k)$. Word2Vec exploits the fact that the masked asset should have an asset embedding that is similar to those of the assets in the context window. We define the average asset embedding in the context window as $\mathbf{x}_i(k) = \frac{\sum_{a \in \mathcal{C}_i(k)} \mathbf{x}_a}{|\mathcal{C}_i(k)|}$. Word2Vec assigns a probability distribution over the masked asset, conditional on the average asset embedding in the context window:

$$\Pr(b_i(k) = a \mid \mathbf{x}_i(k)) = \frac{\exp(\mathbf{x}'_a \mathbf{x}_i(k))}{\sum_{b \notin \mathcal{C}_i(k)} \exp(\mathbf{x}'_b \mathbf{x}_i(k))}. \quad (8)$$

That is, the masked asset should have a high similarity with the average asset embedding in the context window. We train Word2Vec to minimize the cross-entropy loss function over the masked assets:

$$\ell = - \sum_{b_i(k) \in \mathcal{B}} \ln(\Pr(b_i(k) = a_i(k) \mid \mathbf{x}_i(k))). \quad (9)$$

We can modify this methodology to estimate investor embeddings. Instead of ordering the assets for each investor based on its portfolio weights, we rank the investors for each asset based on its ownership shares. Word2Vec assigns a probability that the masked investor ranked k is investor i , conditional on the average investor embedding in the context window. We implement this idea for the transformer model to test its ability to identify similar investors.

Global Vectors for Word Representation (GloVe) is a model that is related to Word2Vec, which is a matrix factorization method based on the co-occurrence matrix of words in a certain context window (Pennington et al., 2014). In Appendix C, we discuss how to use GloVe to estimate asset embeddings.

⁴Alternatively, we could order the assets based on the difference between the portfolio weights and the market weights or a measure of portfolio rebalancing, which we describe in Appendix A.

3.4. Transformer Models

In recommender systems and Word2Vec, the embedding assigned to a word or an asset does not change, depending on the context. For example, the word “bank” has a different meaning in “bank deposit” versus “river bank”. Similarly, Citigroup has a different context as part of a banking portfolio, large-cap portfolio or value portfolio. Following Vaswani et al. (2017), modern LLMs such as BERT and GPT are transformer models with an attention mechanism, which estimate contextualized embeddings. A transformer model can be an encoder that transforms an input sequence into contextualized embeddings or a decoder that transforms contextualized embeddings into an output sequence. We adopt the BERT model, which is a bidirectional encoder.⁵

3.4.1. An Example of the Attention Mechanism

We start with a simple example of the attention mechanism to describe contextualized embeddings, specializing the textbook treatments of this subject to our context (Prince, 2023; Jurafsky and Martin, 2025). Investor i holds a set of assets $\mathcal{A}_i$, and our goal is to estimate the contextualized embedding for asset $a \in \mathcal{A}_i$. We start with an input embedding $\mathbf{x}_a$ for asset a , called the query. We also have an input embedding for $\mathbf{x}_b$ for each asset $b \in \mathcal{A}_i$, called the key. We compute the similarity between the query and the key as $\sigma_{ab} = \mathbf{x}'_a \mathbf{x}_b$. We then compute the contextualized embedding for asset a as a weighted average of input embeddings $\mathbf{x}_b$, called the values:

$$\mathbf{x}_a^i = \frac{\sum_{b \in \mathcal{A}_i} \exp(\sigma_{ab}) \mathbf{x}_b}{\sum_{b \in \mathcal{A}_i} \exp(\sigma_{ab})}. \quad (10)$$

The query $\mathbf{x}_a$ is a high-dimensional vector that captures many aspects of a firm, including its industry classification, reliance on external finance, and supply-chain risk. The contextualized embedding $\mathbf{x}_a^i$ focuses attention on external finance if the other assets in the portfolio have a high similarity through the elements that represent external finance. Clearly, the attention mechanism delivers a different contextualized embedding for asset a , depending on the other assets in the portfolio.

In this simple example, we used the same embeddings for the query, the keys, and the values. Generalizing this example, the transformer model uses $\mathbf{q}_a = \mathbf{W}_q \mathbf{x}_a$ as the query,

⁵The BERT model is a bidirectional encoder, using surrounding words on both sides of the target word. In contrast, the GPT model is a unidirectional decoder, generating the next word based on the previous words. The BERT model is a better fit our application, by estimating contextualized embeddings from the entire portfolio. The GPT model may be suited for some applications such as generative portfolios that we discuss in Section 9.

$\mathbf{k}_b = \mathbf{W}_k \mathbf{x}_b$ as the keys, and $\mathbf{v}_b = \mathbf{W}_v \mathbf{x}_b$ as the values. We need to estimate these matrices $\mathbf{W}$ as part of the training. We compute the similarity between the query and the key as $\sigma_{ab} = \mathbf{q}'_a \mathbf{k}_b$. We then compute the contextualized embedding for asset a as a weighted average of the values:

$$\mathbf{x}_a^i = \frac{\sum_{b \in \mathcal{A}_i} \exp(\sigma_{ab}) \mathbf{v}_b}{\sum_{b \in \mathcal{A}_i} \exp(\sigma_{ab})}. \quad (11)$$

The transformer model stacks multiple attention layers with a feed-forward layer in between. A transformer model is faster and has better performance with multiple attention heads per layer, by allowing the attention heads to focus on different aspects of the portfolio. Michel et al. (2019) provides an in-depth discussion of the benefits of multi-headed attention mechanisms.

What is still missing from the model so far is information about the location of words or, in our case, the position of assets in an investor’s portfolio. Without such information, the model treats the inputs as a bag of words (or bag of stocks). To incorporate this important information, we rank all assets in an investor’s portfolio based on the holdings (or, alternatively, active weights or rebalancing) to determine each asset’s position. We then take the sum of the position embeddings and the input asset embeddings $\mathbf{x}_a$, and use the summed embeddings as inputs into the model. The position embeddings are estimated in modern BERT models alongside the input embeddings and the parameters of the attention layers.⁶

3.4.2. Training the BERT Model

To explain our application of the BERT model, we first summarize the training process for the original BERT model. Researchers train the BERT model on two tasks, masked language modeling and next-sentence prediction. In masked language modeling, we mask one (or multiple) words in a sentence. For example, the BERT model must predict the masked word in “The Fed decided to ____ the target rate to fight inflation.” The BERT model generates a probability distribution over the masked word, based on the other words in the sentence. We estimate the BERT model to maximize the probability of picking the right word on a training sample. To pick the hyper-parameters, we evaluate the model on a validation sample to make sure that the model does not overfit.

The trained BERT model assigns an input embedding $\mathbf{x}_a$ for each word a and a contextualized embedding $\mathbf{x}_a^i$ (see (11)) for each word a in sentence i , which depends on the other words in the sentence. In some applications such as sentence classification or senti-

⁶We also explored a BERT model with sinusoidal position embeddings, following Devlin et al. (2019). However, a BERT model with learned position embeddings performs better in our applications.

ment analysis, we would like to summarize an entire sentence. For this purpose, Reimers and Gurevych (2019) fine-tune the pre-trained BERT model to predict whether two sentences belong together (i.e., two consecutive sentences) or not (i.e., randomly paired sentences). We commonly estimate a sentence embedding by averaging the contextualized embeddings for all words in a sentence as $\frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} \mathbf{x}_a^i$, to which we will refer as the average contextualized embedding. Alternatively, we could use the [CLS] token, which is the first token of a sentence that represents the entire sentence.

We estimate two variants of the BERT model for our application: Portfolio-Shares BERT (PS-BERT) and Ownership-Shares BERT (OS-BERT). In PS-BERT, we use the investor’s portfolio shares to estimate average contextualized investor embeddings, which capture investor similarity. In OS-BERT, we use the asset’s ownership shares to estimate average contextualized asset embeddings, which capture asset similarity.

3.4.3. Portfolio-Shares BERT

To implement PS-BERT, we order the assets in an investor’s portfolio in a decreasing order of portfolio weights. In practice, the input sequence may look like

Tesla, Coinbase Global,..., DraftKings.

We represent the name of each asset as a token, which results in A tokens (i.e., the number of assets) plus a few special tokens.⁷

We estimate the investor embeddings in two steps. In the first step, we train a BERT model with four attention layers, two attention heads per layer, and a context window of 62 assets. We split the portfolio into equal chunks if there are more than 62 assets in the portfolio. In analogy to masked language modeling, we mask one (or more) of the assets in an investor’s portfolio. Following Devlin et al. (2019, Appendix A), we mask 15% of the assets in the training data, of which 80% are assigned the token [MASK], 10% are assigned the actual asset, and 10% are assigned a random asset. We then train the BERT model to predict the masked assets. For example, Figure 1 shows the top ten holdings of the ARK Innovation ETF in July 2023. The BERT model must predict the masked asset in the fourth largest position, which is Zoom Video Communications.

Let $\mathcal{B} = \{b_i(k) = [\text{MASK}]\}$ be the set of all masked assets in the validation data, where the true identity of the masked asset is $b_i(k) = a_i(k)$. The BERT model assigns a probability distribution $\Pr(b_i(k) = a \mid \mathcal{A}_i(k))$ over the masked asset, conditional on the set $\mathcal{A}_i(k)$ of

⁷The special tokens are [CLS] for the beginning of a sentence, [SEP] for the end of a sentence, [UNK] for unknown tokens (which does not apply in our case), [PAD] to complete the length of the sentence to a standard batch size of, in our case, 64, and [MASK] for the masked tokens.

ARKK
ARK Innovation ETF

	Company	Ticker	CUSIP	Shares	Market Value (\$)	Weight (%)
1	TESLA INC	TSLA	88160R101	3,496,872	\$967,024,982.88	12.43%
2	COINBASE GLOBAL INC -CLASS A	COIN	19260Q107	7,945,138	\$620,515,277.80	7.98%
3	ROKU INC	ROKU	77543R102	8,865,426	\$546,110,241.60	7.02%
4	ZOOM VIDEO COMMUNICATIONS-A	ZM	98980L101	8,258,591	\$534,348,251.79	6.87%
5	UIPATH INC - CLASS A	PATH	90364P105	28,152,366	\$463,106,420.70	5.95%
6	BLOCK INC	SQ	852234103	7,069,493	\$456,759,942.73	5.87%
7	EXACT SCIENCES CORP	EXAS	30063P105	4,031,264	\$368,739,718.08	4.74%
8	UNITY SOFTWARE INC	U	91332U101	8,350,868	\$338,627,697.40	4.35%
9	SHOPIFY INC - CLASS A	SHOP	82509L107	5,430,238	\$335,751,615.54	4.32%
10	DRAFTKINGS INC-CL A	DKNG UW	26142V105	12,035,607	\$303,658,364.61	3.90%

Figure 1: An Example of a Masked Portfolio. This figure shows the top ten holdings of the ARK Innovation ETF in July 2023. The fourth largest position, which is Zoom Video Communications, is masked by a red line.

all other assets in the investor’s portfolio. We train the BERT model to minimize the cross-entropy loss function over the masked assets:

$$\ell = - \sum_{b_i(k) \in \mathcal{B}} \ln (\Pr (b_i(k) = a_i(k) \mid \mathcal{A}_i(k))). \quad (12)$$

The trained model assigns an input embedding for each asset and a contextualized embedding for each investor-asset pair through the attention mechanism.

In the second step, we train a sentence transformer model by fine-tuning the pre-trained BERT model. We split the even and odd positions of a portfolio to construct portfolio pairs, where each half contains up to 62 assets. We train the sentence transformer model to predict the portfolio pairs, where the even and odd positions of the same investor are a match. In fine-tuning the sentence transformer, we use cosine similarity between investor embeddings as a distance measure to predict matching pairs. This step ensures that we can use cosine similarity as a distance measure to compare investor embeddings.

With the trained sentence transformer in hand, we compute the average contextualized embedding for a given investor for the largest 62 positions (due to the length of the context window), and refer to it as the investor embedding.

3.4.4. Ownership-Shares BERT

To implement OS-BERT, we order an asset’s investors in a decreasing order of ownership shares. In practice, the input sequence may look like

Vanguard Growth Index Fund, Fidelity Magellan Fund,..., ARK Innovation
ETF.

We represent the name of each investor as a token, which results in I tokens (i.e., the number of investors) plus a few special tokens.

We estimate the asset embeddings for the OS-BERT model, following the same two steps as the PS-BERT model. We first train the BERT model to predict masked investors. The trained model assigns an input embedding to each investor and, for every asset, a contextualized embedding for each investor, which conditions on the other investors holding the asset, through the attention mechanism. In the second step, we train a sentence transformer model by fine-tuning this pre-trained BERT model. We train the sentence transformer model to predict the ownership pairs, where the even and odd positions of the same asset are a match.

We use the trained model to compute the contextualized embedding for each investor holding an asset and refer to the average contextualized embedding for a given asset as the asset embedding. We again average the embeddings of the largest 62 owners due to the length of the context window. Through the fine-tuning step, we make sure that these asset embeddings can be compared using cosine similarity as a distance measure.

In the baseline specification, we train the OS-BERT model on each cross section. In Appendix C, we consider a generalized training procedure that uses a longer sample to pre-train the model and a more recent sample to fine-tune the model. This method uses more data and can provide more accurate estimates of embeddings. However, in our exploration, we found the performance gains to be too limited to add this additional complexity. In the same appendix, we also propose an integrated model of asset and investor embeddings, which we leave for future research.

4. Benchmarks to Evaluate Model Performance

The advancement of AI relies critically on benchmarks to evaluate new models and to compare them against existing models. For example, the image classification benchmarks in the ImageNet Large Scale Visual Recognition Challenge were instrumental in the advancement of deep learning and its application to computer vision. Benchmarks level the playing field for all research teams by establishing clear metrics of progress and success.

We introduce three benchmarks to evaluate the application of AI models to finance. They are predicting relative valuations, explaining the comovement of stock returns, and predicting institutional portfolio decisions. We choose these benchmarks because they satisfy three important criteria. First, the asset pricing model in Section 2 predicts that the portfolio holdings data are informative for these benchmarks. Second, these benchmarks are at the heart of asset pricing and portfolio theory, for which we have existing models (e.g., the Fama-French five-factor model) that establish a high hurdle. Third, and most distinctively, these benchmarks require only a single quarter of data (i.e., a single cross section or a short panel of monthly data). This criterion facilitates a true out-of-sample test to measure progress in real time. For example, we could imagine a quarterly competition to evaluate new models on these benchmarks, just like the ImageNet Large Scale Visual Recognition Challenge. Hellum et al. (2025) analyze such competitions, and propose benchmarks that focus more on the time-series.

4.1. *Relative Valuation Benchmark*

Financial analysts assess the valuation of an existing firm or predict the valuation of a new firm through a comparison with similar firms matched on characteristics (Hommel et al., 2023; Ben-David and Chinco, 2024). The asset pricing model in Section 2 implies that we should define similar firms based on asset embeddings, instead of the usual practice of firm characteristics only. We focus on market-to-book equity as our measure of valuation, but other measures include the price-dividend ratio and the price-earnings ratio.

In Appendix A, we show that log market equity is a linear function of asset embeddings. In each quarter, we first estimate a cross-sectional regression of log market equity on log book equity b_{at} :

$$p_{at} = \gamma_t b_{at} + \alpha_t + p_{at}^\perp. \quad (13)$$

The residual $p_{at}^\perp$ is our valuation measure that is more general than market-to-book equity. Market-to-book equity is a special case when $\gamma_t = 1$.

We randomly split the sample of firms into 80% for training and 20% for out-of-sample testing, keeping the same split for all models. On the training data, we estimate a ridge regression of the valuation on the asset embeddings:

$$p_{at}^\perp = \beta_t' \mathbf{x}_{at} + \delta_t + \epsilon_{at}. \quad (14)$$

In each quarter, we standardize the asset embeddings and estimate the ridge penalty by ten-fold cross validation.⁸ We denote the estimated coefficients as $\hat{\beta}_t$ and $\hat{\delta}_t$.

⁸One potential concern with using cross validation in cross-sectional models is cross-sectional dependence

We evaluate the model performance in the testing data. The performance measure is the out-of-sample R^2 , averaged over the entire sample period

$$\text{RV} = 1 - \frac{1}{T} \sum_{t=1}^T \frac{\text{Var} \left(p_{at}^\perp - \hat{\beta}_t' \mathbf{x}_{at} - \hat{\delta}_t \right)}{\text{Var} \left(p_{at}^\perp \right)}. \quad (15)$$

4.2. Return Comovement Benchmark

A large literature in asset pricing finds variation in average stock returns and comovement of stock returns along firm characteristics. A test of average stock returns requires decades of data, but we could test comovement of stock returns on a single quarter of data. Thus, we test the performance of the asset embeddings against firm characteristics in explaining the comovement of stock returns. Firm characteristics are high hurdle because decades of research have selected those that have the highest explanatory power.

We use the asset embeddings in quarter $t - 1$ to explain the monthly stock returns r_{am} in month m of quarter t . We randomly split the sample of firms into 80% for training and 20% for out-of-sample testing, keeping the same split for all models. On the training data, we estimate a ridge regression of monthly stock returns on the asset embeddings:

$$r_{am} = \beta_m' \mathbf{x}_{a,t-1} + \delta_m + \epsilon_{am}, \quad (16)$$

where the slope β_m is a vector of monthly factor realizations. The constant δ_m removes the common factor with a unit factor loading across stocks. In each quarter, we standardize the asset embeddings and estimate the ridge penalty by ten-fold cross validation within the training sample. We denote the estimated coefficients as $\hat{\beta}_m$ and $\hat{\delta}_m$.

We evaluate the model performance in the testing data. The performance measure is the out-of-sample R^2 , averaged over the entire sample period:

$$\text{RC} = 1 - \frac{1}{M} \sum_{m=1}^M \frac{\text{Var} \left(r_{am} - \hat{\beta}_m' \mathbf{x}_{a,t-1} - \hat{\delta}_m \right)}{\text{Var} \left(r_{am} \right)}. \quad (17)$$

in the error terms, ϵ_{at} . This is not of concern for our benchmarks for two reasons. First, given the micro foundation in Section 2, the cross-sectional dependence is precisely captured by the embeddings, $\mathbf{x}_{at}$. Indeed, the errors in (14) are size-weighted idiosyncratic demand shocks of investors that are cross-sectionally uncorrelated. Second, even if we use lower-dimensional embedding models and some residual correlation in the residuals remains, we are interested in how well embeddings predict valuations (as opposed to causally identifying the impact of specific elements of the embeddings vector on valuations or, in case of the next benchmark, comovement; i.e., we are interested in the R^2 of regression (14), rather than a structural interpretation of coefficient β_t). By design, β_t then captures the projection of valuations (or comovement) on the embeddings space and the residuals are, by construction, uncorrelated. This validates our use of cross validation.

4.3. Managed Portfolio Benchmark

According to the portfolio theory underlying equation (6), investors assign similar portfolio weights to assets with similar asset embeddings. We test this core prediction using the managed portfolios of mutual funds, ETFs, and hedge funds.

We randomly split the sample of investors into 90% for training and 10% for out-of-sample testing, keeping the same split for all models. In the testing data, we mask the k -th largest position in each investor’s portfolio, which we denote as $b_i(k)$. The model must predict the masked asset, based on the set $\mathcal{A}_i(k)$ of all other assets in the investor’s portfolio. We set $k = 2$ in our implementation.⁹

In the training data, we estimate the recommender system, Word2Vec, and the PS-BERT model. To predict the masked asset, the model must assign a probability distribution over the masked asset. As we discussed in Section 3, Word2Vec and the PS-BERT model directly assign a probability distribution as part of the training process. For the recommender system, we start with equation (6) and make an additional assumption that ϵ_{ia} is drawn from a logistic distribution. Then the probability distribution over the masked asset is

$$\Pr(b_i(k) = a \mid \mathcal{A}_i(k)) = \frac{\exp(\boldsymbol{\lambda}'_i \mathbf{x}_a + \delta_a)}{\sum_{b \notin \mathcal{A}_i(k)} \exp(\boldsymbol{\lambda}'_i \mathbf{x}_b + \delta_b)}. \quad (18)$$

Intuitively, the masked asset should be the largest position among the set of assets that the investor does not hold, if the investor were to hold all assets.

We evaluate the model performance in the testing data. Let $\mathcal{B} = \{b_i(k) = [\text{MASK}]\}$ be the set of all masked assets in the testing data, where the true identity of the masked asset is $b_i(k) = a_i(k)$. The performance measure is the average log likelihood of correct predictions across investors, relative to the log likelihood of random guesses (i.e., $\frac{1}{|\mathcal{A}_i(k)|}$):

$$\text{MP} = \frac{\sum_{b_i(k) \in \mathcal{B}} \ln(\Pr(b_i(k) = a_i(k) \mid \mathcal{A}_i(k)))}{-|\mathcal{B}| \ln(|\mathcal{A}_i(k)|)} - 1. \quad (19)$$

A higher measure implies better performance with an upper bound of one.

4.4. Proceeding Pragmatically without Statistical Guarantees

Following much of the AI and ML literature, we proceed without statistical guarantees (i.e., unbiasedness, consistency, or optimality). Researchers have not yet developed the statistical tools to rigorously analyze the consistency and optimality of Word2Vec and the

⁹We could increase the sample size by varying the position of the masked asset (i.e., $k = 2, 3, \dots$). However, it is not necessary for our implementation because we already have a large sample of investors.

transformer models. Indeed, no one would claim that LLMs are optimal. Instead, researchers simply observe that these models perform well in many empirical applications. We proceed pragmatically by adopting models that have proven useful in other AI and ML applications and testing them on our benchmarks.

Researchers have developed asymptotic theory for the PCA model. In particular, Bai (2003) provides sufficient conditions for identification of factor models in the limit of large cross-sectional and time dimensions. These results apply to our setting in the limit of infinitely many investors and assets. The asymptotic theory also implies consistency in the relative valuation benchmark, as we have discussed above. We do not devote space to repeating the results that are valid only for the PCA model to keep a unified treatment of all models in Section 3. Instead, we refer interested readers to the relevant literature (Bai, 2009; Fortin et al., 2026).

5. Data

We provide a summary of the data construction with further details in Appendix D. We also describe competing methods to characterize firms, including firm characteristics and the text-based embeddings from Cohere and OpenAI.

5.1. Data Construction

Our data construction essentially follows Kojen et al. (2024). We use the stock holdings of mutual funds, ETFs, closed-end funds, variable annuity funds, and hedge funds from the FactSet Ownership Data (FactSet, 2024). The data for hedge funds are from the Form 13F filings. For the other types of investors, we use data at the fund level instead of the 13F filer level (i.e., Vanguard funds instead of the Vanguard Group) for a more granular view of portfolios.¹⁰ Future research may be able gain a more granular and complete view of institutional and households portfolios by bringing in additional data sources (e.g., Gabaix et al., 2023). Custodial data, which cover a large number of investors at a high frequency, would be ideal for this purpose.

We use the data on firm characteristics and stock returns from Jensen et al. (2023), which are based on the CRSP US Stock Database (Center for Research in Security Prices, 2024) and Compustat Fundamentals (S&P Global, 2024). We aggregate the stock-level data to the firm level by the CRSP permco. We remove micro and nano caps as defined by Jensen et al. (2023). We keep stocks that are held by at least 20 investors and investors who hold

¹⁰In an earlier version of this paper, we used only the institution-level data from the Form 13F filings. The asset embeddings performed well on our benchmarks, but not as well as using the fund-level data. We expect this result because more granular portfolio holdings lead to better estimates of the asset embeddings.

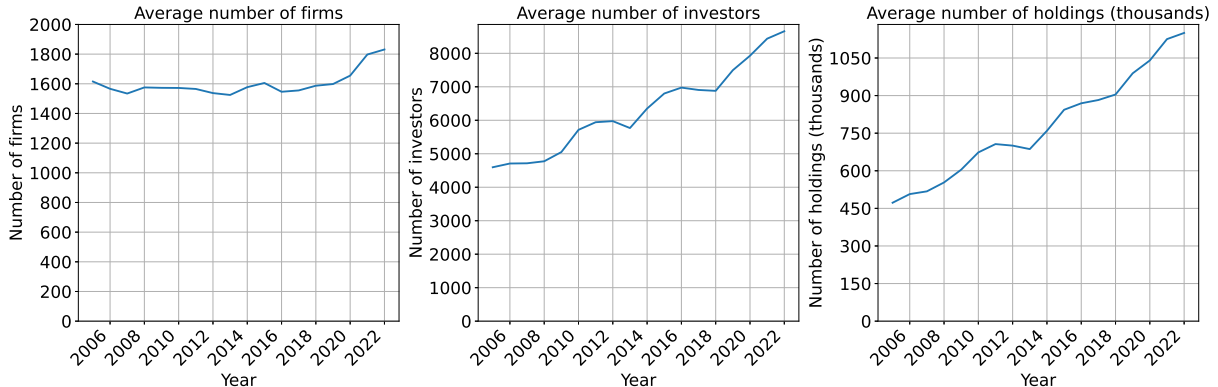


Figure 2. Number of Firms, Investors, and Stock Holdings. The quarterly observations for 2005.Q1 to 2022.Q4 are averaged by year.

at least 20 stocks. In Section 8, we use cleaned transcripts of earnings calls from FactSet to interpret the asset embeddings. The quarterly sample period is 2005.Q1 to 2022.Q4.

Figure 2 shows the time series of the number of firms (left panel), investors (middle panel), and stock holdings (right panel). The number of funds and stock holdings increase substantially over time with the growing importance of institutional investors. In the context of LLMs, a larger and more diverse corpus enables researchers to estimate a larger transformer model for a fixed vocabulary (Kaplan et al., 2020). Thus, the growing number of stock holdings could enable researchers to estimate larger models of asset embeddings in the future.¹¹

5.2. Competing Methods to Characterize Firms

We use the benchmarks in Section 4 to evaluate the relative performance of three methods to characterize firms. They are holdings-based embeddings, firm characteristics, and text-based embeddings from Cohere and OpenAI.

The first method to characterize firms is holdings-based embeddings. For each method in Section 3, we estimate embeddings whose dimension match the number of firm characteristics that we describe below. We also estimate a larger model with ten dimensions, which is relatively small in the context of AI. Therefore, we also estimate larger models up to 128 dimensions for the recommender systems and the transformer models.

Observed firm characteristics are a natural competitor to characterize a firm. For the relative valuation benchmark, we use four characteristics that include the market beta, asset growth, profitability, and the dividend-asset ratio. For this benchmark, we cannot use price-

¹¹An analysis of the scaling laws for transformer models based on portfolio holdings data is an interesting topic for future research.

based characteristics such as market-to-book equity and momentum that have a mechanical correlation with the outcome variable. For the return comovement and managed portfolio benchmarks, we use six characteristics that include the market beta, log market equity, log market-to-book equity, asset growth, profitability, and momentum. We limit our sample to firms for which these characteristics are not missing.

The choice of firm characteristics always requires discretion. We choose the five characteristics in the Fama and French (2015) model and momentum as a high hurdle for the relative valuation and return comovement benchmarks. Decades of research have selected these characteristics for their high explanatory power. Consequently, these characteristics are among the most prominent in the asset pricing literature. Since our benchmark comparisons hold the dimension of the asset embeddings and the firm characteristics constant at six, we do not consider larger models of firm characteristics. However, a larger model of firm characteristics may be productive for some practical applications (Bryzgalova et al., 2025; Jensen et al., 2023).

The third method to characterize firms is text-based embeddings from two leading AI labs, Cohere and OpenAI. We use the most recent publicly available model as of March 2024. These embeddings represent not only the name of a firm (e.g., Apple as a fruit) but also its business (e.g., Apple as a technology firm). For example, OpenAI’s ChatGPT model gives a detailed description of a firm’s business strategies and activities in response to a user prompt. As we describe in Appendix D, we reduce the dimension of the embeddings to four or ten for the relative valuation benchmark.

The text-based embeddings are current snapshots, which we cannot use for historical backtesting due to a look-ahead bias (Sarkar and Vafa, 2024). Therefore, we test the text-based embeddings on a single cross section at the end of our sample period. Cohere and OpenAI estimate the text-based embeddings on vast amounts of text data for use in a broad set of applications. We could fine-tune the text-based embeddings to improve performance on a specific benchmark (Sarkar, 2025). Similarly, we could fine-tune the holdings-based embeddings in Section 3, which we leave for future research.

For the remainder of the paper, we use the labels in Table 1 to refer to the competing models of firms and investors. We refer back to Section 3 for a detailed description of these models.

6. Comparing Methods to Characterize Firms

The result in Section 2 shows that all features relevant for asset prices are embedded in portfolio holdings. This prompts two questions. First, how well do observed characteristics

Table 1. Competing Models of Firms and Investors

Label	Description
Placebo	Random embeddings drawn from a standard normal distribution with the same dimension as the base model of asset embeddings.
Beta	Only the market beta as a characteristic.
Characteristics	Relative valuation benchmark: Market beta, asset growth, profitability, and dividend-asset ratio. Return comovement and managed portfolio benchmarks: Market beta, log market equity, log market-to-book equity, asset growth, profitability, and momentum.
RS-Binary	Recommender system that uses only binary information for positive versus zero positions.
RS-Ranks	Recommender system that uses only percentile ranks and replaces the zero positions with zeros.
RS-L-0	Recommender system that uses log dollar holdings, removes the investor and asset fixed effects, and replaces the zero positions with zeros.
RS-L-Min	Recommender system that uses log dollar holdings and replaces the zero positions with the investor’s smallest positive position.
RS-L	Recommender system that uses log dollar holdings and ignores zero positions.
Word2Vec	Word2Vec estimated on stocks ordered by the investor’s portfolio weights.
PS-BERT	Transformer model trained on the masked portfolio shares of investors.
OS-BERT	Transformer model trained on the masked ownership shares of stocks.

For all models of asset embeddings, the dimension of the base model is the number of firm characteristics in the corresponding benchmark. The large model has ten dimensions.

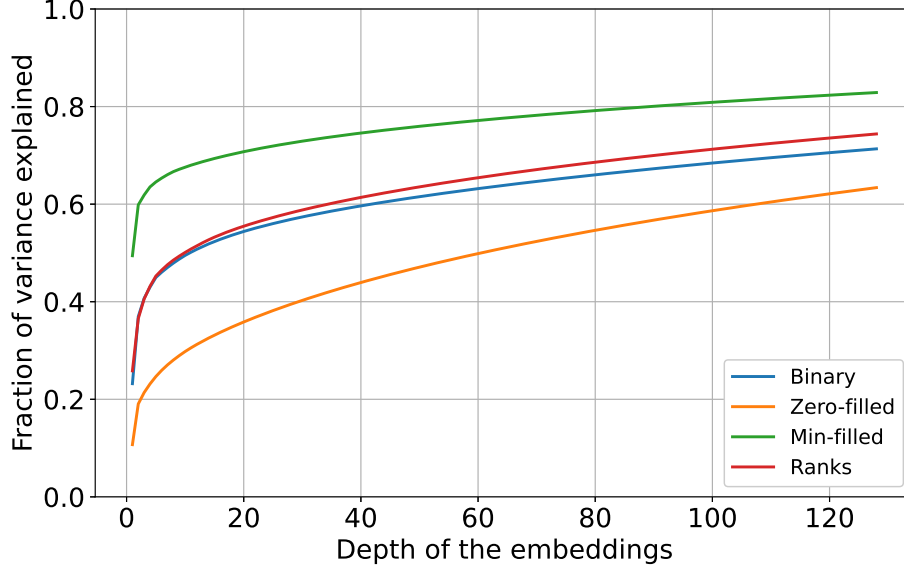


Figure 3. Variation in the Portfolio Holdings Explained by the Recommender Systems. The models are RS-Binary, RS-Ranks, RS-L-0, and RS-L-Min. See Table 1 for the model descriptions. The embedding dimension ranges from 4 to 128. The quarterly estimates for 2005.Q1 to 2022.Q4 are averaged over the entire sample period.

and text-based embeddings explain holdings data? Second, firm characteristics and, perhaps in particular, text data may contain information that is not typically used by investors. Again following the result in Section 2, this information should then also be irrelevant for asset prices. We therefore explore whether holdings-based embeddings explain variation in observed characteristics and text-based embeddings. In the analysis, as a point of reference, we use the holdings-based embeddings estimated using the recommender system methodology.

6.1. The factor structure in portfolio holdings

We first examine how well the recommender systems in Section 3 explain the cross section of portfolio holdings. For each recommender system that we estimate using PCA, we compute the share of the cross-sectional variance of the portfolio holdings that the asset embeddings explain. We then average this statistic over the entire sample period. Figure 3 reports the share of the cross-sectional variance of the portfolio holdings as a function of the embedding dimension from four to 128.

We have three main findings. First, all models explain most of the cross-sectional variance of the portfolio holdings when the embedding dimension is sufficiently high. When the embedding dimension is 128, the explained variation ranges from 63% for the RS-L-0 model to 83% for the RS-L-Min model. Second, the explained variation increases sharply for the

first few dimensions, implying a strong factor structure in the portfolio holdings. Third, there are meaningful differences in explanatory power across the models. The RS-L-Min model has the highest explanatory power for a given embedding dimension, followed by the RS-Ranks model. These models ensure a smooth transition from the smallest positive position to the zero positions. The RS-Binary and RS-Ranks models have similar explanatory power with only small differences at higher dimensions.

6.2. Comparing methods to characterize firms

We now turn to the two main questions in this section. First, we explore how well different methods to characterize firms can explain holdings data. The previous section shows that holdings-based embeddings explain the majority of the variation in holdings, and we now explore how other methods fare to explain holdings data. For observed characteristics¹² and the text-based embeddings from Cohere and OpenAI, we extract the first K principal components.

For each of the four methods to characterize firms, we then regress each investor’s holdings on the K principal components and measure the fraction of the variance that can be explained. We use the holdings converted to percentile ranks as our holdings data for this exercise, to mitigate any impact of outliers. Formally, we estimate the coefficients b_{ik} and c_i ($i = 1 \dots I$, $k = 1 \dots K$) in:

$$h_{ia} = \sum_{k=1}^K b_{ik} X_{ak} + c_i + e_{ia}. \quad (20)$$

We measure the performance of a method to characterize firms as

$$R^2 = 1 - \frac{\sum_{i,a} e_{ia}^2}{\sum_{i,a} (h_{ia} - \bar{h}_i)^2}, \quad (21)$$

where $\bar{h}_i := \frac{1}{A} \sum_a h_{ia}$.

Panel A of Figure 4 presents the results. In relative terms, observed characteristics explain more of the variation in portfolio holdings than text-based embeddings from either Cohere or OpenAI. However, there is a significant difference in explanatory power relative to holdings-based embeddings (which provide an upper bound). As a point of reference, we add randomly simulated embeddings (Placebo Embeddings), which demonstrates that all embedding measures have some meaningful explanatory power. That said, the main takeaway is that there is important common variation in portfolio holdings that is not well captured by

¹²For the JKP characteristics (Jensen et al. (2023)), we first convert each characteristics to percentile ranks, normalized to $[-1, 1]$.

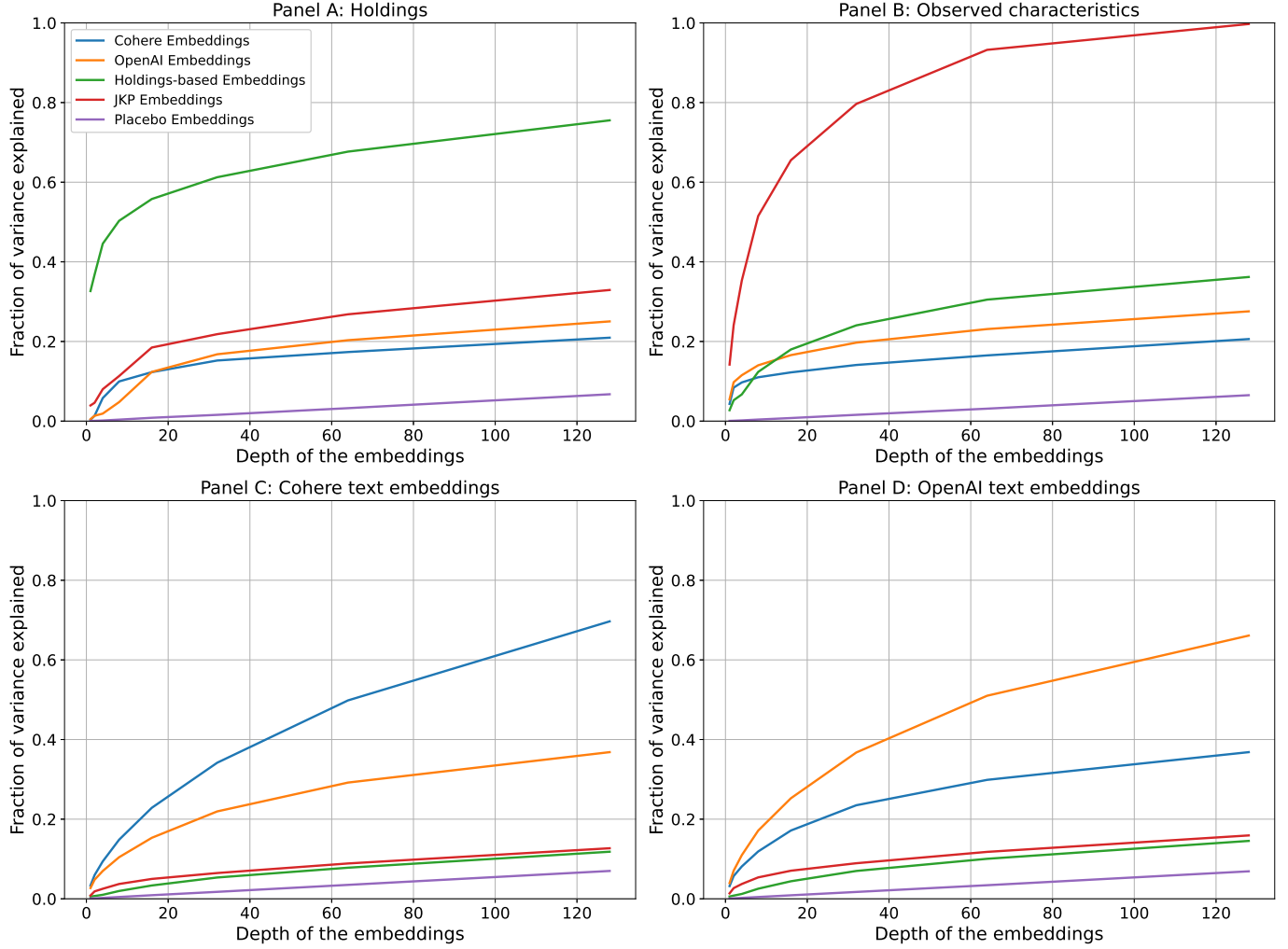


Figure 4. Variation in Portfolio Holdings Explained by Observed Characteristics and Text-based Embeddings We estimate the fraction of variation in portfolio holdings (Panel A), 153 observed characteristics (Panel B), text-based embeddings from Cohere (Panel C), and text-based embeddings from OpenAI (Panel D) that can be explained by asset embeddings or principal components (embeddings) estimated from observed characteristics, text-based embeddings from Cohere or text-based embeddings from OpenAI. The depth of the embeddings varies from 1 to 128.

either observed characteristics or text-based embeddings. As text-based embeddings depend on the text that is being embedded (in addition to the embedding model), this analysis can be extended using other text data (e.g., from earnings calls or 10-K filings). Put broadly, the methodology in Panel A can be used to assess which text-based information is reflected in investors' portfolios and, by market clearing, potentially in prices as well.

Next, we explore whether observed characteristics and text-based embeddings can be explained by the other methods to characterize firms. As before, we take holdings-based embeddings, the principal components of observed characteristics, and principal components of text-based embeddings from either Cohere or OpenAI as methods to characterize firms. We start from the data matrix for observed characteristics and either set of text-based embeddings. We denote the data matrix by $D \in \mathbb{R}^{A \times M}$, where $M = 153$ for observed characteristics, $M = 1,024$ for the Cohere embeddings, and $M = 3,072$ for the OpenAI embeddings. We then estimate coefficients b_{mk} and c_m ($m = 1 \dots M$, $k = 1 \dots K$) in:

$$D_{ma} = \sum_{k=1}^K b_{mk} X_{ak} + c_m + e_{ma}, \quad (22)$$

and compute a measure of fit

$$R^2 = 1 - \frac{\sum_{m,a} e_{ma}^2}{\sum_{m,a} (D_{ma} - \bar{D}_m)^2}, \quad (23)$$

where $\bar{D}_m := \frac{1}{A} \sum_a D_{ma}$.

The results are presented in Panel B - D of Figure 4. By construction, the own principal components perform best to explain either observed characteristics or text-based embeddings. Other data typically cannot explain the variation in observed characteristics or text embeddings. Interestingly, text embeddings from Cohere best explain the variation in text embeddings from OpenAI and vice versa, but there is a significant amount of independent variation as well. Thus, even conditional on the text, different embedding models generate a nontrivial amount of independent variation. Observed characteristics and holdings-based embeddings cannot explain the variation in text-based embeddings well.

7. Main Empirical Results

We evaluate the performance of asset embeddings, firm characteristics, and text-based embeddings from Cohere and OpenAI on the benchmarks in Section 4.

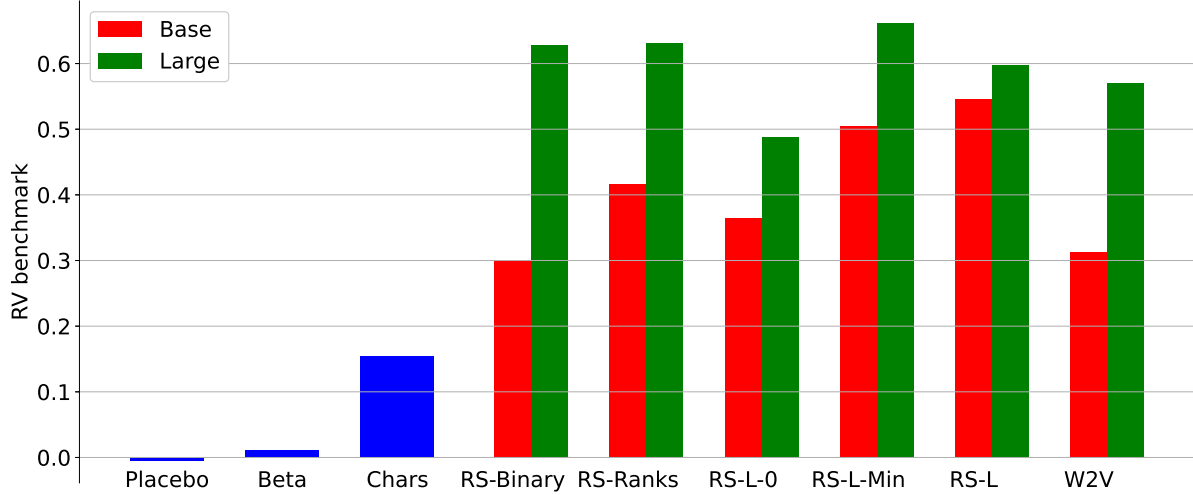


Figure 5. Relative Valuation Benchmark. From left to right, the models are placebo embeddings; the market beta as a characteristic; four characteristics (i.e., market beta, asset growth, profitability, and dividend-asset ratio); recommender systems RS-Binary, RS-Ranks, RS-L-0, RS-L-Min, and RS-L; and Word2Vec. See Table 1 for the model descriptions. The base model has four dimensions, and the large model has ten dimensions. The quarterly sample period is 2005.Q1 to 2022.Q4.

7.1. Relative Valuation Benchmark

Figure 5 reports the performance of the competing models in the relative valuation benchmark, which we described in Section 4. The firm characteristics are the market beta, asset growth, profitability, and the dividend-asset ratio. For each model of asset embeddings, we consider a base model with four dimensions to match the firm characteristics and a larger model with ten dimensions. We evaluate the models based on the performance measure (15), which is an out-of-sample R^2 , over the entire sample period.

As we expect, the placebo embeddings have no explanatory power for relative valuations. The market beta has little explanatory power for relative valuations, but the four characteristics explain about 15% of relative valuations. The base models of asset embeddings with four dimensions significantly outperform the four characteristics, and the larger models with ten dimensions further improve the performance. Interestingly, both the RS-Binary model that uses only the extensive margin and the RS-L model that uses only the intensive margin benefit from a larger dimension (Table 1 describes the models). Among the models of asset embeddings, the RS-L-Min model has consistently high performance at both dimensions.

In Figure 6, the first bar repeats the performance of the large model of asset embeddings in Figure 5. The second bar reports the performance of a combined model of asset embeddings and the four characteristics. The performance gain from adding the four characteristics is

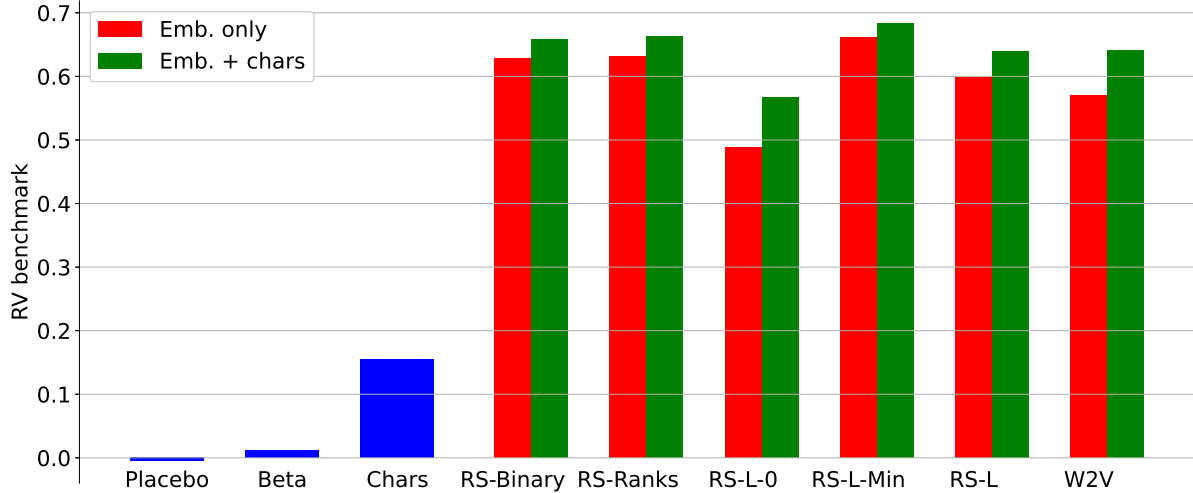


Figure 6. Combining Asset Embeddings and Firm Characteristics in the Relative Valuation Benchmark. The first bar repeats the performance of the large model of asset embeddings in Figure 5. The second bar reports the performance of a combined model of the asset embeddings and the four characteristics (i.e., market beta, asset growth, profitability, and dividend-asset ratio). The quarterly estimation sample covers 2005.Q1 to 2022.Q4.

small, implying that the asset embeddings already contain much of the information in these characteristics.

One might think that this success of asset embeddings at explaining valuations is in part due to the fact that some investors chose to have holdings proportional to market capitalization. We think that this is not the case, for at least two reasons. First, equation (6) models the active deviations of the portfolio weights from the market weights, given its firm fixed effect δ_a . Second, RS-Binary and RS-Ranks ignore dollar magnitudes altogether, relying only extensive margins or ranks.

We compare the performance of holdings-based embeddings with that of text-based embeddings from Cohere and OpenAI. As we described in Section 5, we reduce the dimensions of the text-based embeddings to match the base (four-dimensional) and large (ten-dimensional) models of asset embeddings. Since the text-based embeddings are current snapshots, we limit the sample period to a single cross section in 2022.Q4. Figure 7 shows that the text-based embeddings have little explanatory power for relative valuations.

To understand the relatively poor performance of the text-based embeddings, Table 2 reports the ten closest firms to Apple, Citigroup, and Walmart, using the text-based embeddings from OpenAI. Although these embeddings identify firms with similar businesses, semantic similarity appears to also play an important role. For example, some of the firms that are closest to Citigroup are banks, but other firms like Cigna and Caci International

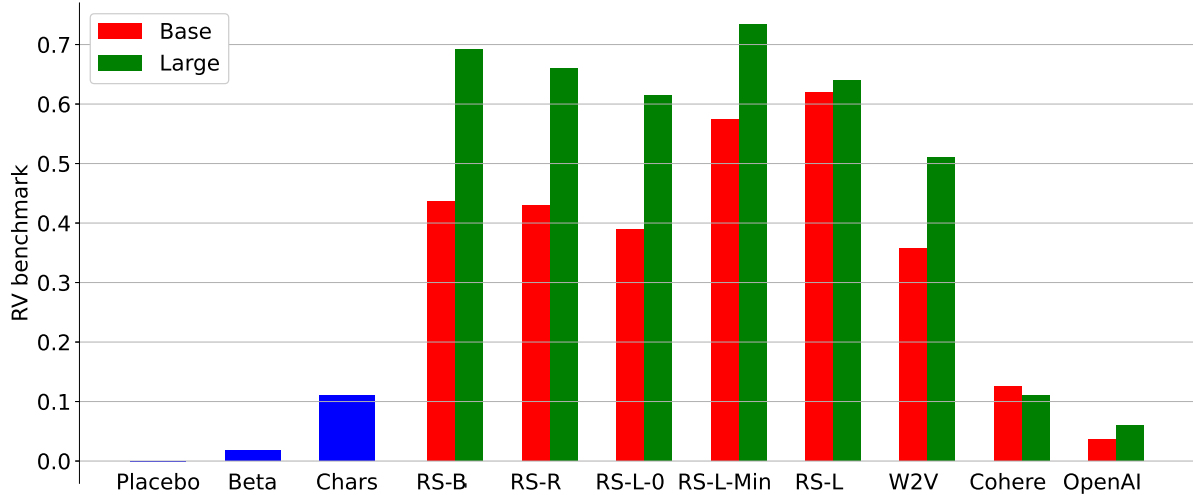


Figure 7. Text-Based Embeddings in the Relative Valuation Benchmark. The dimension of the text-based embeddings from Cohere and OpenAI are reduced to match the base (four-dimensional) and large (ten-dimensional) models of asset embeddings. The sample period is 2022.Q4.

simply start with “C” or “Ci”. Similarly, the closest firm to Apple is Appian and to Walmart is Walgreens. As we discussed in Section 5, we could fine-tune the text-based embeddings to potentially improve performance, which we leave for future research.

Figure 8 compares the performance of the RS-L-Min model and the OS-BERT model at high dimensions up to 128. We choose the RS-L-Min model as the best performing model at ten dimensions. Although the OS-BERT model has high explanatory power for relative valuations, the RS-L-Min model outperforms at all dimensions. The performance of the OS-BERT model is more sensitive to the embedding dimension than that of the RS-L-Min model. The performance of the RS-L-Min model starts to flatten at 16 dimensions, but the performance of the OS-BERT model keeps increasing beyond 16 dimensions.

These results suggest that the OS-BERT model may benefit from fine-tuning on the relative valuation benchmark. Indeed, we typically train LLMs on masked language modeling and subsequently fine-tune them on a specific task. We leave fine-tuning for future research, given the already solid performance of the OS-BERT model across the benchmarks in Section 4.

7.2. Return Comovement Benchmark

Figure 9 reports the performance of the competing models in the return comovement benchmark, which we described in Section 4. The firm characteristics are the market beta, log market equity, log market-to-book equity, asset growth, profitability, and momentum. For

Table 2. Text-Based Embeddings from OpenAI

Rank	Apple Inc	Citigroup Inc	Walmart Inc
1	Appian Corp	Citizens Financial Group Inc	Walgreens Boots Alliance Inc
2	Adobe Inc	Goldman Sachs Group Inc	Home Depot Inc
3	Interdigital Inc	American International Group Inc	Murphy USA Inc
4	Microsoft Corp	Comerica Inc	Amazon Com Inc
5	Gopro Inc	Cigna Corp New	Qurate Retail Inc
6	Netapp Inc	Capital One Financial Corp	Big Lots Inc
7	Intel Corp	Caci International Inc	Burlington Stores Inc
8	Alphabet Inc	Capital City Bank Group	Dollar Tree Inc
9	Autodesk Inc	C N O Financial Group Inc	Nordstrom Inc
10	Appfolio Inc	JP Morgan Chase & Co	Kohls Corp

This table reports the ten closest firms to Apple, Citigroup, and Walmart, according to the text-based embeddings from OpenAI.

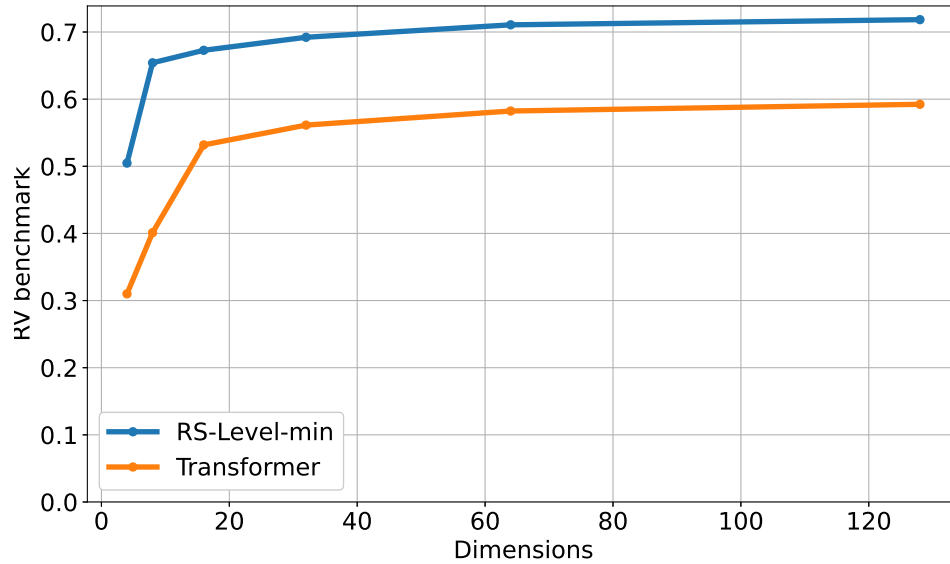


Figure 8. High-Dimensional Models in the Relative Valuation Benchmark. This figure reports the performance of the RS-L-Min model and the OS-BERT model at dimensions four to 128. The quarterly sample period is 2005.Q1 to 2022.Q4.

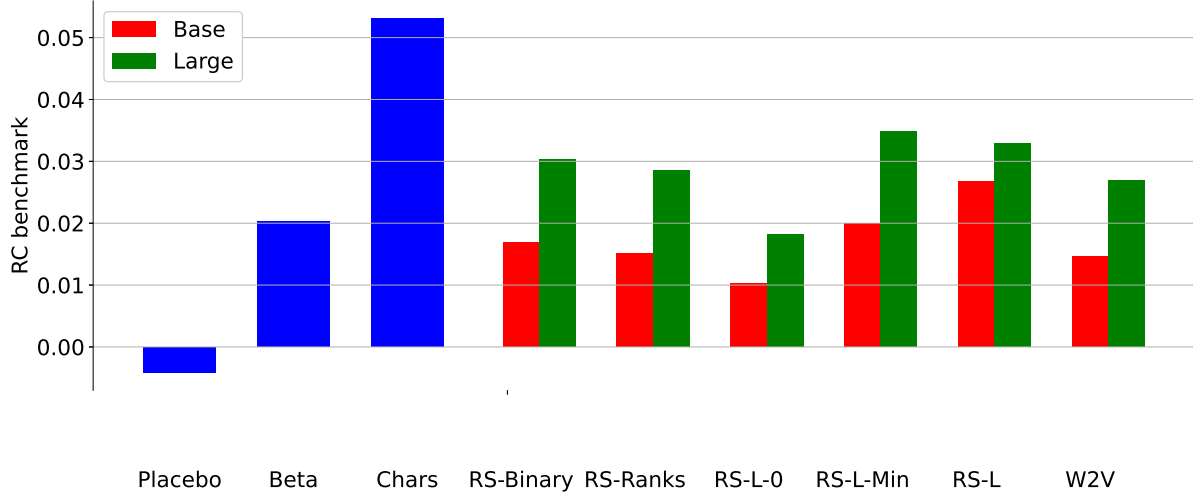


Figure 9. Return Comovement Benchmark. From left to right, the models are placebo embeddings; the market beta as a characteristic; six characteristics (i.e., market beta, log market equity, log market-to-book equity, asset growth, profitability, and momentum); recommender systems RS-Binary, RS-Ranks, RS-L-0, RS-L-Min, and RS-L; and Word2Vec. See Table 1 for the model descriptions. The base model has six dimensions, and the large model has ten dimensions. The asset embeddings or firm characteristics in quarter $t - 1$ are used to explain the monthly stock returns in quarter t . The quarterly sample period is 2005.Q1 to 2022.Q4.

each model of asset embeddings, we consider a base model of six dimensions to match the firm characteristics and a larger model with ten dimensions. We evaluate the models based on the performance measure (17), which is an out-of-sample R^2 , over the entire sample period.

As we expect, the placebo embeddings have no explanatory power for stock returns. The market beta explains 2% of the variation in stock returns, and the six characteristics explain 5.3%. These characteristics are a high hurdle because decades of research have selected them to explain the comovement of stock returns. The large models of asset embeddings with ten dimensions outperform the market beta but fall short of the high hurdle of six characteristics. Among the large models of asset embeddings, the RS-L-Min model has the highest performance.

In Figure 10, the first bar repeats the performance of the large model of asset embeddings in Figure 9. The second bar reports the performance of a combined model of asset embeddings and the six characteristics. All models outperform a baseline of the six characteristics only, implying that the asset embeddings contain independent information that helps to explain the comovement of stock returns.

Figure 11 compares the performance of the RS-L-Min model and the OS-BERT model

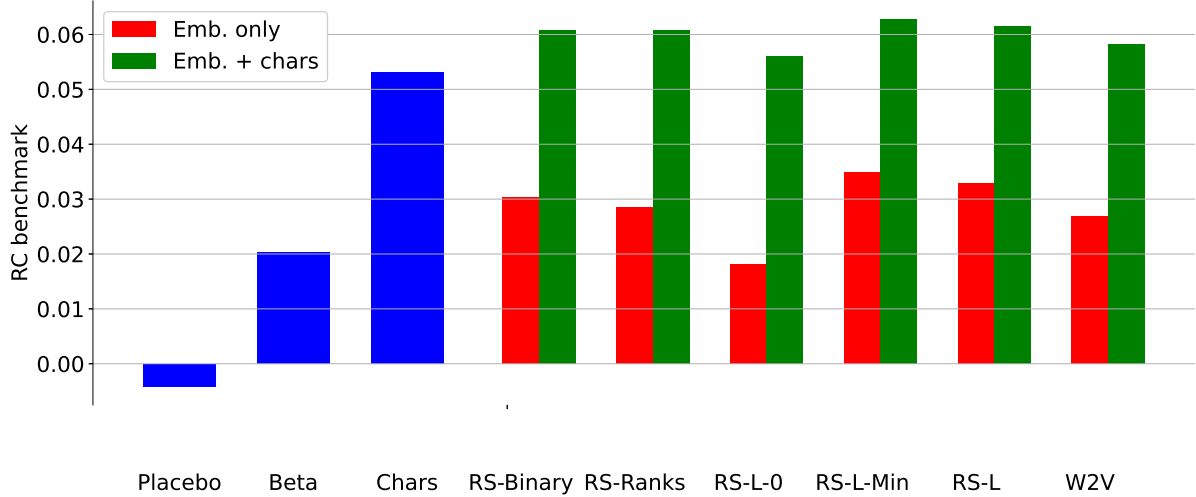


Figure 10. Combining Asset Embeddings and Firm Characteristics in the Return Comovement Benchmark. The first bar repeats the performance of the large model of asset embeddings in Figure 9. The second bar reports the performance of a combined model of the asset embeddings and the six characteristics (i.e., market beta, log market equity, log market-to-book equity, asset growth, profitability, and momentum). The asset embeddings or firm characteristics in quarter $t - 1$ are used to explain the monthly stock returns in quarter t . The quarterly sample period is 2005.Q1 to 2022.Q4.

at high dimensions up to 128. We choose the RS-L-Min model as the best performing model at ten dimensions. The performance of the RS-L-Min model peaks with an out-of-sample R^2 of 7.2%, exceeding 5.3% for the six characteristics. Thus, the higher dimensions of asset embeddings contain information that helps to explain the comovement of stock returns. The OS-BERT model does not perform as well as the RS-L-Min model, but its performance continues to improve at high dimensions.

7.3. Managed Portfolio Benchmark

Figure 12 reports the performance of the competing models in the managed portfolio benchmark, which we described in Section 4. The firm characteristics are the market beta, log market equity, log market-to-book equity, asset growth, profitability, and momentum. For each model of asset embeddings, we consider a base model with six dimensions to match the firm characteristics and a larger model with ten dimensions. We evaluate the models based on the performance measure (19) over the entire sample period.

In contrast to the previous two benchmarks, holdings-based embeddings estimated using the Word2Vec model significantly outperform firm characteristics and holdings-based embeddings estimated using recommender systems. Word2Vec has a performance measure of 26% for the base model and 29% for the large model. The firm characteristics and the

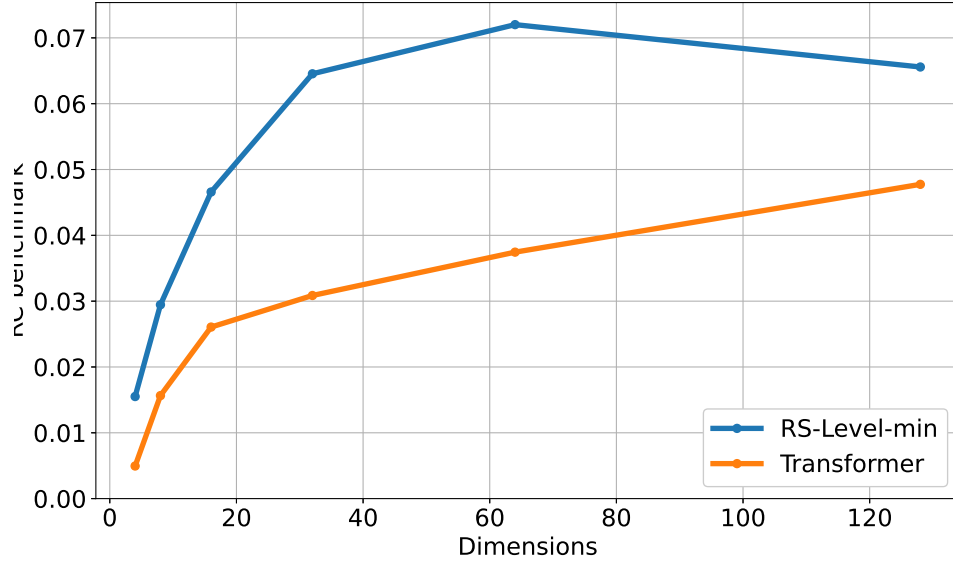


Figure 11. High-Dimensional Models in the Return Comovement Benchmark. This figure reports the performance of RS-L-Min recommender system and the OS-BERT model at dimensions four through 128. The asset embeddings in quarter $t - 1$ are used to explain the monthly stock returns in quarter t . The quarterly sample period is 2005.Q1 to 2022.Q4.

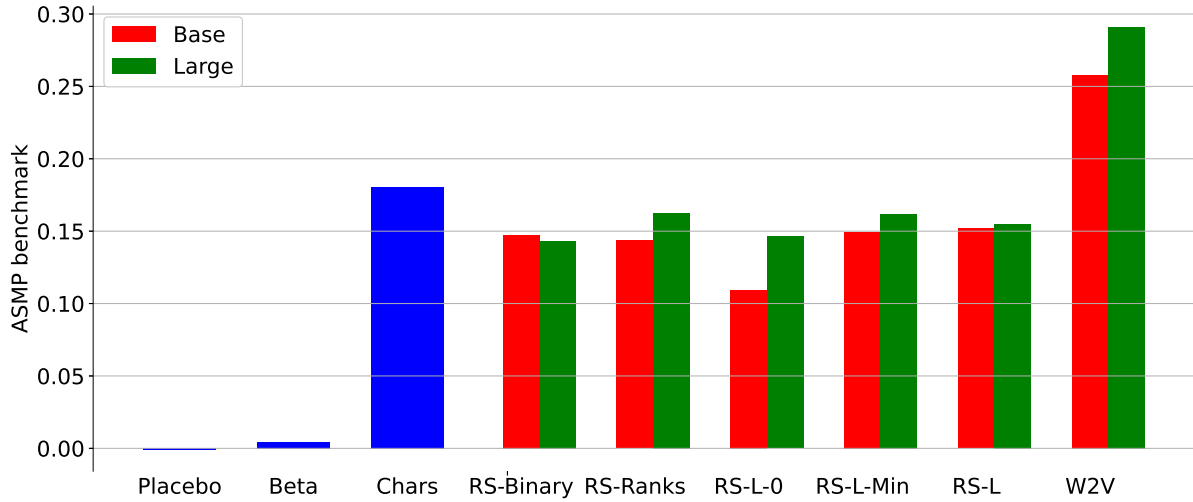


Figure 12. Managed Portfolio Benchmark. From left to right, the models are placebo embeddings; the market beta as a characteristic; six characteristics (i.e., market beta, log market equity, log market-to-book equity, asset growth, profitability, and momentum); recommender systems RS-Binary, RS-Ranks, RS-L-0, RS-L-Min, and RS-L; and Word2Vec. See Table 1 for the model descriptions. The base model has six dimensions, and the large model has ten dimensions. The sample period is 2022.Q4.

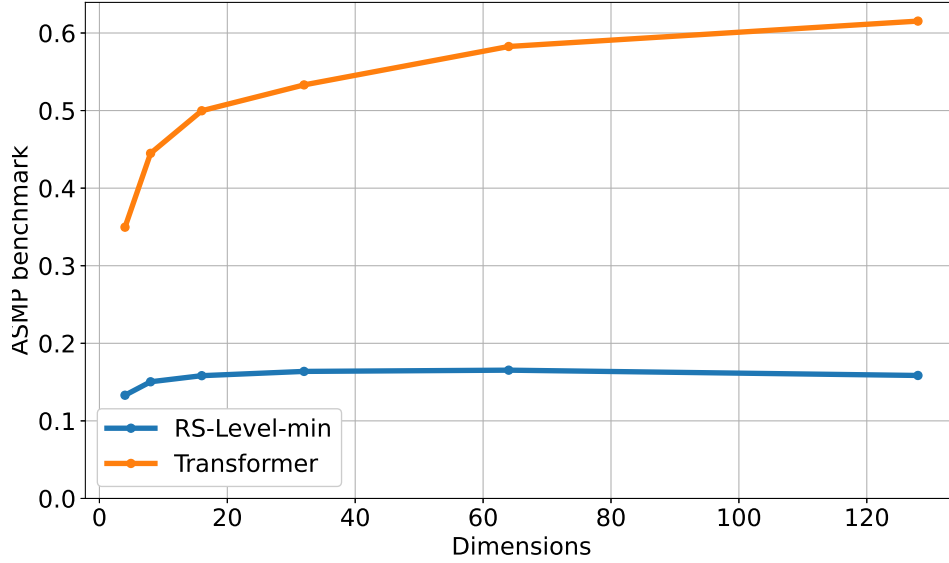


Figure 13. High-Dimensional Models in the Managed Portfolio Benchmark. This figure reports the performance of RS-L-Min recommender system and the PS-BERT model at dimensions four to 128. The sample period is 2022.Q4.

recommender systems have a performance measure below 20%.

Figure 13 compares the performance of the RS-L-Min model and the PS-BERT model at high dimensions up to 128. We train the PS-BERT model on masked assets in managed portfolios, which is precisely the objective of this benchmark. Even at low dimensions, the PS-BERT model outperforms Word2Vec and all other models. The performance measure for the PS-BERT model increases from 35% at four dimensions to over 60% at higher dimensions. The performance of the RS-L-Min model does not increase much at higher dimensions.

We offer three possible explanations for the large performance gap between the RS-L-Min model and the PS-BERT model. First, the RS-L-Min model is bilinear in asset and investor embeddings, while the PS-BERT model is nonlinear. The PS-BERT model may capture nonlinearities that are important for the managed portfolio benchmark. Second, the RS-L-Min model requires an auxiliary assumption that the probability distribution over the masked asset is a logit function (18), so an alternative specification may improve its performance. Third, the PS-BERT model minimizes a cross-entropy loss function over the masked assets, while the RS-L-Min model minimizes a least-squares loss function for log dollar holdings. The RS-L-Min model may outperform because the loss function directly aligns with the managed portfolio benchmark.

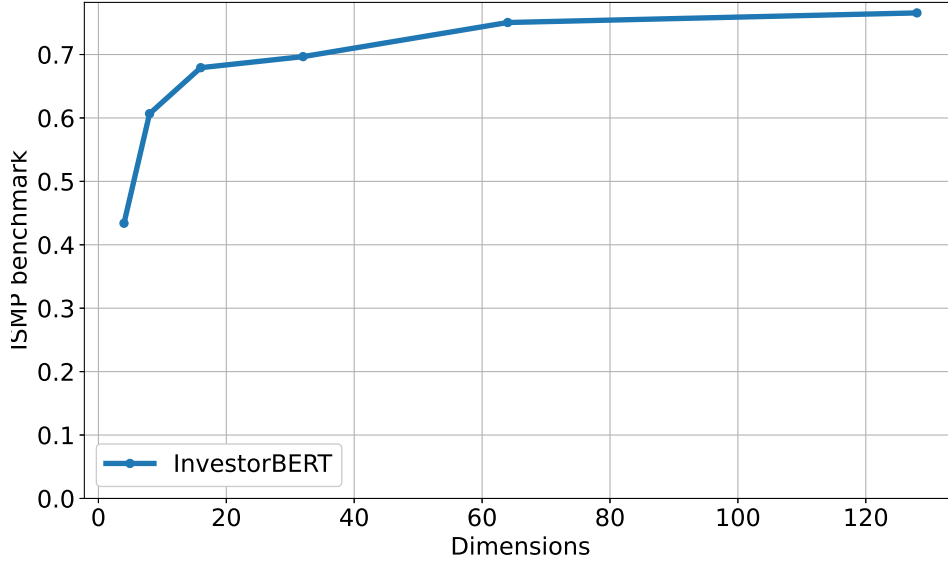


Figure 14. Masked Investor Benchmark. This figure reports the performance of the OS-BERT model at dimensions four through 128. The sample period is 2022.Q4.

7.4. Masked Investor Benchmark

In analogy to the managed portfolio benchmark, we define a masked investor benchmark to test the performance of the OS-BERT model. Instead of predicting the second largest asset in a given portfolio, we predict the second largest investor for a given asset. The investor embeddings could identify similar investors beyond standard investor characteristics such as institutional type (e.g., mutual fund, hedge fund, or insurance company), size, benchmark portfolio, and measures of activeness (e.g., turnover and active share).

Figure 14 reports the performance of the OS-BERT model at high dimensions up to 128. The OS-BERT performs well in the masked investor benchmark, just as the PS-BERT model performs well on the managed portfolio benchmark in Figure 13. The performance measure for the OS-BERT model increases from 43% at four dimensions to over 70% at higher dimensions.

8. Interpreting Embeddings

The previous section showed that asset and investor embeddings are latent vector representations that perform well on the new benchmarks that we introduce. This naturally raises the question what the embeddings capture and how assets and investors are represented economically. While traditionally it is hard to interpret latent embeddings, we propose a simple methodology to provide interpretability in our context.

To start, we list similar firms and investors using holdings-based asset and investor em-

beddings in Section 8.1. We then provide economic narratives to understand why these assets (Section 8.2) and investors (Section 8.3) share economically meaningful similarities. For the textual analysis, we use the Gemini 2.5 Pro model from Google, which is a state-of-the-art reasoning model as of June 2025. While we can list only a handful of examples in the paper, we provide a companion website (<https://www.mktgen.ai>) that allows readers to explore embeddings and their economic interpretation, based on the latest cross-section of regulatory filings. We also use the latest model in the Gemini series to process the text data on the companion website.

8.1. Using Embeddings to Identify Similar Assets and Investors

The first two columns of Table 3 report the nine closest firms to Apple and Citigroup in 2022.Q4, based on the average contextualized embeddings from the OS-BERT model. At first glance, much of the model output conforms to our intuition for asset similarity. For example, Tesla, Amazon, Microsoft, Alphabet (Google), and Nvidia are technology firms that share similarities with Apple. Likewise, American International Group, Wells Fargo, Goldman Sachs, and Bank of America are financial firms that share similarities with Citigroup. However, industry is only one of many characteristics that investors consider in portfolio choice. Investors may also attend to other characteristics such as profitability, riskiness, or ESG score. The multi-dimensional nature of asset similarity could easily explain why the model output does not always align along industry characteristics. In Section 8.2, we use text data to provide a richer economic narrative that goes beyond simple industry groupings.

The last two columns of Table 3 report the nine closest investors to the Dimensional US Small Cap Value ETF and AQR Arbitrage in 2022.Q4, based on the average contextualized embeddings from the PS-BERT model. These funds have very different investment strategies. The Dimensional US Small Cap Value ETF follows a more traditional strategy that is familiar from the asset pricing literature (Fama and French, 1992) and Morningstar’s classification system for funds. AQR Arbitrage follows a less traditional strategy. FactSet summarizes AQR Arbitrage’s investment strategy as “specializes in global merger arbitrage, global convertible arbitrage and other event driven strategies. The firm also manages a limited number of alternative investment strategies, including special purpose acquisition companies (SPACs), closed-end fund arbitrage, distressed securities, and private investment in public equity (PIPEs). AQR employs arbitrage investment analysis to analyze various risks and return characteristics of an investment. They also utilize fundamental securities analysis to implement its arbitrage strategies and other alternative investment strategies.”

The model output conforms to our intuition for investor similarity. For the DFA ETF, eight of the nine closest investors have “small cap value” in their name. The lone exception

Table 3. Asset and Investor Similarity in the Transformer Models

Apple Inc	Citigroup Inc	Dimensional US Small Cap Value ETF	AQR Arbitrage LLC
Tesla Inc	Altria Group Inc	Acclivity Small Cap Value	BCK Capital Management LP
Costco Wholesale Corp	Exxon Mobil Corp	Undiscovered Managers Behavioral Value	Water Island Capital LLC
Amazon Com Inc	American International Group Inc	SEI Inst. Managed Trust - Small Cap Value	VIA AM SICAV - Alternative-Liquid
Microsoft Corp	Wells Fargo & Co New	SBL Series Q (Small Cap Value)	GAMCO International SICAV - Merger Arbitrage
Nike Inc	General Motors Co	Guggenheim Small Cap Value	Yakira Capital Management, Inc.
Alphabet Inc	Valero Energy Corp New	MassMutual Small Company Value	GDL
Nvidia Corp	Gilead Sciences Inc	MML Small Company Value	Pentwater Capital Management LP
Adobe Inc	Goldman Sachs Group Inc	PF Small Cap Value	Lyxor Newcits IRL - Tiedemann Arbitrage Strategy
Disney Walt Co	Bank Of America Corp	MML Small/Mid Cap Value	Gabelli & Co. Investment Advisers, Inc.

We use the average contextualized embeddings from the OS-BERT model to identify the nine closest firms to Apple and Citigroup. We use the average contextualized embeddings from the PS-BERT to identify the nine closest investors to the Dimensional US Small Cap Value ETF and AQR Arbitrage. We define the closest firms and investors based on cosine similarity of the 128-dimensional contextualized embeddings. The sample period is 2022.Q4.

is Undiscovered Managers Behavioral Value which is in fact also a small-cap value fund, marketed as “Designed to provide long-term capital appreciation primarily through a portfolio of small-cap value stocks.” We reemphasize that the input into the PS-BERT model is portfolio holdings data and not the investor names or textual descriptions of the strategies.

For AQR Arbitrage, the names of the closest investors do not directly reveal similarity, except for a few cases with “arbitrage strategy” or “merger arbitrage” in the name. We therefore turn to text data in Section 8.3 to provide a richer narrative of the similarities in investment strategies.

8.2. *Interpreting Asset Embeddings*

We feed the earnings calls that took place during the first quarter of 2023 to the LLM. We first prompt the LLM to produce a summary for each of the earnings call transcripts:

Analyze this earnings call transcript and provide a structured summary providing:

****Investor Appeal Factors****

- What makes this company attractive for portfolio inclusion
- Type of company (e.g. technology, financial services, etc.)
- Growth drivers and expansion opportunities
- Market trends the company is positioned to benefit from
- Risks to which the company is exposed and potential headwinds

Focus on extracting insights that would influence portfolio construction decisions rather than just reporting numbers. Don’t mention company names or products. Refer to the company as "the company" and the products as "the products".

We remove company and product names to limit the model’s reliance on information in its training set, although such knowledge may be a virtue if our focus is on interpretability and developing economic narratives. We then feed all summaries to the LLM using the prompt (in addition to formatting instructions):

Below are summaries of 10 earnings transcripts of companies that appear often together in the portfolios of investors. Provide the common investor appeal factors of the companies.

This analysis yields the following results.

Understanding the commonality in 10 firms most similar to Apple

- **Market Leadership and Dominant Positioning:** The companies are consistently described as leaders in their respective industries, possessing dominant market share, strong brand power, and valuable intellectual property or franchises.
- **Positioned for Major Secular Growth Trends:** A primary appeal is their strategic positioning to capitalize on durable, long-term market shifts. Artificial Intelligence (AI) is a near-universal theme, cited as a core growth driver and a new computing platform. Other key trends include the enterprise shift to cloud computing, digital transformation, and the growth of the digital/creator economy.
- **Financial Strength and Resilience:** They demonstrate robust financial health, often reporting record revenues, strong operating margins, and significant free cash flow generation, even in challenging macroeconomic environments.
- **Strong Customer Loyalty and Ecosystems:** Many of the companies have built powerful, sticky ecosystems that foster high customer loyalty and recurring revenue. This is evident through massive installed user bases, membership models with high renewal rates, and direct-to-consumer strategies.
- **Clear Growth and Innovation Strategies:** Each company has well-defined avenues for future growth, including aggressive investment in product innovation, expansion into new product categories or services, and a focus on international market penetration.
- **Commitment to Shareholder Returns:** A common factor is a stated commitment to returning capital to shareholders, frequently through significant share repurchase programs and, for some, the payment or reinstatement of dividends.

Understanding the commonality in 10 firms most similar to Citigroup

- **Strong Shareholder Returns:** Nearly all companies emphasize a strong commitment to returning capital to shareholders through consistent dividends (often with a history of increases) and significant share repurchase programs.
- **Robust Financial Performance and Profitability:** A common theme is the delivery of strong financial results, including record or industry-leading earnings, robust cash flow, and expanding profit margins in core businesses.
- **Market Leadership:** The companies are frequently leaders in their respective industries or in highly profitable market segments, which provides a stable financial foundation and competitive advantage.
- **Strong Balance Sheet and Capital Position:** A key appeal is maintaining a strong or "fortress" balance sheet, characterized by high capital levels (e.g., CET1), low debt, and financial flexibility to navigate economic cycles and fund growth.

- **Clear Strategic Initiatives for Growth:** Each company presents a clear strategic plan for future growth, whether through transformation (e.g., pivoting to EVs, digital banking, or smoke-free products), operational turnarounds, strategic simplification, or investment in new high-growth areas like oncology and low-carbon solutions.
- **Positioned for Favorable Market Trends:** The companies are strategically positioned to capitalize on major secular and cyclical trends, such as rising interest rates, the energy transition, digital adoption, and shifts in consumer preferences.
- **Operational Excellence and Cost Discipline:** Many highlight successful cost-reduction programs, operational efficiency gains, and disciplined management as key drivers for improving profitability and offsetting headwinds like inflation.

The summaries of the earnings call identify themes relevant to investors, such as profitability, growth prospects, and corporate policies. While broad categories are naturally related, the details differ meaningfully for Apple and Citigroup. For example, firms related to Apple have strong brand value and operating performance, while firms related to Citigroup have strong balance sheets and cost efficiency. Depending on the question of interest, the summaries, and the aggregation of summaries, can be adjusted and refined by modifying the prompts. In our experience, this approach works well to identify coherent economic narratives that underlie the asset embeddings.

8.3. *Interpreting Investor Embeddings*

We use a similar methodology to interpret investor embeddings. For U.S. mutual funds and ETFs, we use the descriptions of the investment strategies in summary prospectuses (Abis et al., 2022). Specifically, Form 497K contains a section “Principal Investment Strategies”, which describes the strategy in detail. For hedge funds, we use the description from FactSet (which is typically a couple of sentences). For international mutual funds, we still rely on the prospectus or, when unavailable, we use the strategy description on a fund’s website.

We feed the summaries to the LLM with the prompt (in addition to formatting instructions):

Below are descriptions of investment strategies of different funds. Identify the *common* investment styles and approaches how they choose stocks.

Additional instructions:

- Do not include boilerplate text.
- Focus on common components across the descriptions.
- Be concise.
- Only use information from the descriptions.

- Focus on the core of the strategy.

We provide some additional guidance as prospectuses contain significant amounts of boilerplate language (e.g., the fund invests 80% of its AUM in equities, policies regarding derivatives and short positions, et cetera). This analysis yields the following responses.

Understanding the commonality in 10 investors most similar to Dimensional US Small Cap Value ETF

- **Focus on Small-Capitalization Companies:** All funds target U.S. companies with small market capitalizations, often using the Russell 2000 Index or similar benchmarks as a reference for the size range.
- **Value Investing Philosophy:** The core strategy is to identify and invest in stocks that are considered "undervalued," "attractively valued," or have a "low relative price."
- **Use of Specific Valuation Metrics:** Value is assessed by comparing a stock's price to its fundamental business measures. Common metrics cited include price-to-book value, price-to-earnings, and price-to-cash-flow.
- **Blend of Quantitative and Fundamental Analysis:** The funds frequently combine quantitative, factor-based models to screen for potential investments with traditional fundamental research into a company's financial health and business model.
- **Consideration of Quality and Other Factors:** In addition to value, many funds incorporate other factors to refine their selections, such as company profitability (e.g., return on equity) and momentum.

Understanding the commonality in 10 investors most similar to AQR Arbitrage

- **Event-Driven Investing:** The strategies primarily focus on investing in securities of companies undergoing specific corporate events, such as mergers, acquisitions, reorganizations, liquidations, spin-offs, and other special situations.
- **Arbitrage:** A core approach is "arbitrage," specifically seeking to profit from the price discrepancies that arise during these corporate events (e.g., merger arbitrage).
- **Fundamental Analysis:** Most funds rely on fundamental, bottom-up research to evaluate the viability, risks, and potential outcomes of each investment opportunity.

- **Long/Short Positions:** A common technique is to take both long and short positions, often to hedge against broader market risk and isolate the return from the specific corporate event.

These summaries capture the essence of the trading strategies and provide additional context beyond the names of the funds. For example, instead of thinking of traditional, academic small-cap value strategies, where value is measured by a firm’s market-to-book ratio, funds related to the Dimensional small-cap value ETFs use a broad set of valuation measures, which are at times enriched by quality measures. In terms of methodology, they combine traditional and quantitative measures. Investors similar to the AQR Arbitrage fund focus on corporate events. In doing so, they hedge out aggregate risk factors as much as possible to isolate the (perceived) arbitrage opportunities.

Taken together, these analyses show how to use text data to provide economic narratives to interpret holdings-based embeddings that identify similar assets and investors. This approach overcomes the black-box nature of holdings-based embeddings and enables a deeper economic analysis of asset and investor similarity.

9. Additional Applications of Embeddings

We discuss additional applications of asset embeddings, including generative portfolios, risk management and stress testing, and firm valuation. We also discuss additional applications of investor embeddings, including investor classification, performance evaluation, and detecting crowded trades. Finally, we discuss alternative data sources to maximize explanatory power.

9.1. *Generative Portfolios*

We could use asset embeddings to construct replicating portfolios without historical data on stock returns, which is especially useful for new opportunities and risks not reflected in historical returns. For example, suppose that we would like to construct a COVID-19 factor that has low returns when COVID-19 infections increase. A starting point for this strategy could be long United Airlines and short Zoom, which are exemplar stocks for exposure to COVID-19 risks. To construct a diversified COVID-19 factor, we could use the asset embeddings to find stocks that are most similar to United Airlines and Zoom. We could use the same logic to construct an AI factor, starting from large players such as Alphabet, Amazon, Apple, Meta, Microsoft, and Nvidia.

9.2. Risk Management and Stress Testing

Financial institutions, regulators, and central banks could use asset and investor embeddings for risk management and stress testing. For example, stable diffusion is a generative AI model that creates new images from text descriptions (Rombach et al., 2022). Similarly, regulators and central banks would like to generate stress scenarios that are possible but never realized in the past, which satisfy realistic substitution patterns across assets and investors. They could generate counterfactual changes in asset and investor embeddings, which imply counterfactual changes in asset prices in an asset demand system (Kojien and Yogo, 2019). Thus, we propose a novel methodology to generate stress scenarios, which could have broad application in risk management.

9.3. Firm Valuation

Security analysts have long used the cost of capital implied by the CAPM to value a firm, given a firm’s expected cash flows. Although faculty teach this basic approach to undergraduates and MBA students, they are aware that the CAPM is a simplified model that does not explain actual valuations (Chicago Booth, 2023). Even multi-factor models have poor predictions for actual valuations (Hommel et al., 2023).

An alternative approach is to estimate the cost of capital based on similar firms. However, a systematic definition of firm similarity is elusive. For example, is Apple more similar to technology or consumer-product firms? The asset pricing model in Section 2 implies that the asset embeddings directly reveal how investors perceive similarity. Confirming this theory, asset embeddings predict relative valuations better than firm characteristics in Section 4.

We could estimate the cost of capital based on asset embeddings in two ways. First, we could use the present-value identity to infer the cost of capital, based on the firm’s expected cash flows and the valuation predicted by the asset embedding. Second, we could estimate the cost of capital as the predicted value from a predictive regression of stock returns on the asset embeddings.

9.4. Identifying Peer Investors and Performance Measurement

Investor embeddings identify investors who implement similar investment strategies. Therefore, we could use the investor embeddings to systematically classify investors based on their investment strategies. Standard investor characteristics include institutional type (e.g., mutual fund, hedge fund, or insurance company), size, and measures of activeness (e.g., turnover and active share). However, these investor characteristics do not explain much of the variation in the investment strategies (Kojien et al., 2024).

In this context, investor embeddings can be used for performance evaluation. Daniel et al. (1997) develop characteristics-based benchmarks, based on market equity, market-to-book equity, and momentum. We could extend their insight to higher dimensions, which is important in light of the more recent literature cited in the introduction. We define a set of investors that are similar to investor i , based on cosine similarity of the investor embeddings, as $\mathcal{I}_i = \left\{ j \mid \frac{\lambda_i' \lambda_j}{\|\lambda_i\| \|\lambda_j\|} \geq \chi \right\}$, where $\chi > 0$ is some threshold. We then define investor i 's abnormal return as its return minus the average return of similar investors:

$$\alpha_{it} = r_{it} - \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} r_{jt}. \quad (24)$$

9.5. Crowded Trades

As investor embeddings identify common factors in investment strategies, it is a natural measure of crowded trades. This is especially true when no “valuation ratios” help detect it, as the case for momentum (Stein (2009)). Examples of crowded trades that caused market dislocations include quantitative equity strategies in August 2007 and the Japanese carry trade in August 2024. Early detection of crowded trades, based on the investor embeddings, should be valuable to investors and regulators.

9.6. Alternative Data Sources

The literature has explored many sources of text data, including 10-K filings (Hoberg and Phillips, 2016; Chen and Sarkar, 2020), earnings calls (Hassan et al., 2019), and news articles (Binsbergen et al., 2023; Bybee et al., 2023; Sarkar, 2025). These text data could supplement the portfolio holdings data. An obvious extension of our research is to estimate high-dimensional embeddings based on portfolio holdings data, a large set of firm characteristics, and a collection of the text data to maximize explanatory power. We could also estimate lower-dimensional embeddings through fine-tuning or supervised estimation on a target task.

10. Conclusion

We made a five-fold contribution centered around the concept of holding-based embeddings: asset embeddings to represent firms and investor embeddings to represent investors and their strategies. We show theoretically that portfolio holdings contain all relevant information about the cross-section of asset prices, and are ideal for estimating asset and investor embeddings. This follows directly from the fact that asset prices are generally determined

by an asset demand system and market clearing. The mapping from holdings to embeddings depends on the details of the demand system, and we therefore use various methods to estimate embeddings, building on insights from the recent NLP literature. Indeed, just as documents structure words in ways that allow us estimate word embeddings, portfolios structure assets in ways that allow us to estimate asset embeddings. Thus, this paper provides a new connection between recent advances in AI and ML to applications in economics and finance.

We confirm empirically that asset and investor embeddings make powerful predictions in economics and finance applications, using three new benchmarks that we propose in this paper. A key feature of our benchmarks is that they can be evaluated using a single cross-section of data, allowing for quick iteration and improvements.

Our main empirical finding is that asset embeddings predict relative valuations, explain the comovement of stock returns, and predict the portfolio choice of institutional investors. For example, a base model of four-dimensional asset embeddings explains over 50% of the variation in relative valuations, whereas a set of four leading firm characteristics explains only 15%. Across all benchmarks, we find success training high-dimensional asset embeddings by ridge regression with cross validation. As a point of reference, we compare our embeddings to traditional characteristics and text-based embeddings, and we find that holdings-based embeddings perform well.

Thus, an obvious extension is to estimate high-dimensional asset embeddings on a combination of portfolio holdings data, a large set of firm characteristics, and a collection of text data to maximize explanatory power. Alternatively, fine-tuning or supervised estimation may achieve similar (or greater) explanatory power with a lower-dimensional model. Lastly, we show how text data from firms and investors can be used to provide economic interpretability of asset and investor embeddings, respectively.

We focused on the US equity market as the application in this paper. It would be interesting to explore if our methodology is useful in other countries and asset classes such as fixed income, currencies, commodities, and derivatives. The key input is security-level portfolio holdings data for institutional investors or households (Koijen et al., 2021; Gabaix et al., 2023). In an application to fixed income markets, Gabaix et al. (2025) find that firm embeddings explain credit spreads and the volatility of credit spreads better than credit ratings and the distance to default.

References

- Abis, Simona, Andrea M. Buffa, Apoorva Javadekar, and Anton Lines**, “Learning from Prospectuses,” 2022. Working paper.
- Ali, Usman and David Hirshleifer**, “Shared Analyst Coverage: Unifying Momentum Spillover Effects,” *Journal of Financial Economics*, 2020, *136* (3), 649–675.
- Bai, Jushan**, “Inferential Theory for Factor Models of Large Dimensions,” *Econometrica*, 2003, *71* (1), 135–171.
- Bai, Jushan**, “Panel Data Models with Interactive Fixed Effects,” *Econometrica*, 2009, *77* (4), 1229–1279.
- Balasubramaniam, Vimal, John Y. Campbell, Tarun Ramadorai, and Benjamin Ranish**, “Who Owns What? A Factor Model for Direct Stockholding,” *Journal of Finance*, 2023, *78* (3), 1545–1591.
- Balestriero, Randall and Yann LeCun**, “LeJEPa: Provable and Scalable Self-Supervised Learning without the Heuristics,” 2025. Working paper.
- Beckmeyer, Heiner and Timo Wiedemann**, “Recovering Missing Firm Characteristics with Attention-Based Machine Learning,” 2023. Working paper.
- Ben-David, Itzhak and Alex Chincó**, “Expected EPS $\times$ Trailing P/E,” 2024. NBER Working Paper 32942.
- Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin**, “A Neural Probabilistic Language Model,” *Journal of Machine Learning Research*, 2003, *3*, 1137–1155.
- Betermier, Sebastien, Laurent E. Calvet, Samuli Knüpfer, and Jens Soerlie Kvaerner**, “Investor Factors,” *Journal of Finance*, 2025, *80* (5), 2789–2830.
- Binsbergen, Jules H. van, Svetlana Bryzgalova, Mayukh Mukhopadhyay, and Varun Sharma**, “(Almost) 200 Years of News-Based Economic Sentiment,” 2023. NBER Working Paper 32026.
- Bryzgalova, Svetlana, Sven Lerner, Martin Lettau, and Markus Pelger**, “Missing Financial Data,” *Review of Financial Studies*, 2025, *38* (3), 803–882.

- Bryzgalova, Svetlana, Victor DeMiguel, Sicong Li, and Markus Pelger**, “Asset-Pricing Factors with Economic Targets,” 2023. Working paper.
- Bybee, Leland, Bryan Kelly, and Yinan Su**, “Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text,” *Review of Financial Studies*, 2023, *36* (12), 4759–4787.
- Center for Research in Security Prices**, *US Stock Database* 2024.
- Chen, Andrew Y. and Jack McCoy**, “Missing Values Handling for Machine Learning Portfolios,” *Journal of Financial Economics*, 2024, *155* (103815), 1–15.
- Chen, Jiafeng and Suproteem K. Sarkar**, “A Semantic Approach to Financial Fundamentals,” 2020. Working paper.
- Chen, Luyang, Markus Pelger, and Jason Zhu**, “Deep Learning in Asset Pricing,” *Management Science*, 2024, *70* (2), 714–750.
- Chicago Booth**, “Discount Rates,” 2023. <https://www.kentclarkcenter.org/surveys/discount-rates/>.
- Clayton, Christopher and Antonio Coppola**, “Financial Regulation and AI: A Faustian Bargain?,” 2025. Working paper.
- Cohen, Lauren and Andrea Frazzini**, “Economic Links and Predictable Returns,” *Journal of Finance*, 2008, *63* (4), 1977–2011.
- Cong, Lin William, Ke Tang, Jingyuan Wang, and Yang Zhang**, “AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI,” 2022. Working paper.
- Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers**, “Measuring Mutual Fund Performance with Characteristic-Based Benchmarks,” *Journal of Finance*, 1997, *52* (3), 1035–1058.
- DeMiguel, Victor, Alberto Martin-Utrera, Francisco J. Nogales, and Raman Uppal**, “A Transaction-Cost Perspective on the Multitude of Firm Characteristics,” *Review of Financial Studies*, 2020, *33* (5), 2180–2222.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova**, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” in Jill Burstein, Christy Doran, and Thamar Solorio, eds., *Proceedings of the 2019 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 Association for Computational Linguistics Minneapolis, MN 2019, pp. 4171–4186.

Dolphin, Rian, Barry Smyth, and Ruihai Dong, “Stock Embeddings: Learning Distributed Representations for Financial Assets,” 2022. Working paper.

Dumais, Susan T., George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman, “Using Latent Semantic Analysis to Improve Access to Textual Information,” in Elliot Soloway, Douglas Frye, and Sylvia B. Sheppard, eds., *Proceedings of the ACM CHI 88 Human Factors in Computing Systems Conference*, ACM Press Washington, DC 1988, pp. 281–285.

FactSet, *Ownership Data* 2024.

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

Fama, Eugene F. and Kenneth R. French, “A Five-Factor Asset Pricing Model,” *Journal of Financial Economics*, 2015, 116 (1), 1–22.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, “Taming the Factor Zoo: A Test of New Factors,” *Journal of Finance*, 2020, 75 (3), 1327–1370.

Fortin, Alain-Philippe, Patrick Gagliardini, and Olivier Scaillet, “Latent Factor Analysis in Short Panels,” *Journal of Econometrics*, 2026, 255 (106249), 1–18.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, “Dissecting Characteristics Nonparametrically,” *Review of Financial Studies*, 2020, 33 (5), 2326–2377.

Freyberger, Joachim, Bjorn Hoppner, Andreas Neuhierl, and Michael Weber, “Missing Data in Asset Pricing Panels,” 2022. Working paper.

Gabaix, Xavier and Ralph S. J. Koijen, “In Search of the Origins of Financial Fluctuations: The Inelastic Markets Hypothesis,” 2023. Working paper.

Gabaix, Xavier, Ralph S. J. Koijen, Federico Mainardi, Sangmin Oh, and Motohiro Yogo, “Asset Demand of U.S. Households,” 2023. NBER Working Paper 32001.

Gabaix, Xavier, Ralph S. J. Koijen, Robert J. Richmond, and Motohiro Yogo, “Upgrading Credit Pricing and Risk Assessment through Embeddings,” 2025. Working paper.

- Gu, Shihao, Bryan Kelly, and Dacheng Xiu**, “Empirical Asset Pricing via Machine Learning,” *Review of Financial Studies*, 2020, 33 (5), 2223–2273.
- Hansen, Lars Peter and Scott F. Richard**, “The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models,” *Econometrica*, 1987, 55 (3), 587–613.
- Hassan, Tarek A., Stephan Hollander, Laurence Van Lent, and Ahmed Tahoun**, “Firm-Level Political Risk: Measurement and Effects,” *Quarterly Journal of Economics*, 2019, 134 (4), 2135–2202.
- Hellum, Oliver, Theis Ingerslev Jensen, Bryan T. Kelly, and Lasse Heje Pedersen**, “The Power of the Common Task Framework,” 2025. Working paper.
- Hoberg, Gerard and Gordon Phillips**, “Text-Based Network Industries and Endogenous Product Differentiation,” *Journal of Political Economy*, 2016, 124 (5), 1423–1465.
- Hommel, Nicolas, Augustin Landier, and David Thesmar**, “Corporate Valuation: An Empirical Comparison of Discounting Methods,” 2023. Working paper.
- Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen**, “Is There a Replication Crisis in Finance?,” *Journal of Finance*, 2023, 78 (5), 2465–2518.
- Jurafsky, Dan and James H. Martin**, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. 2025.
- Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei**, “Scaling Laws for Neural Language Models,” 2020. Working paper.
- Kelly, Bryan and Dacheng Xiu**, “Financial Machine Learning,” *Foundations and Trends in Finance*, 2023, 13 (3-4), 205–363.
- Kelly, Bryan T., Seth Pruitt, and Yinan Su**, “Characteristics Are Covariances: A Unified Model of Risk and Return,” *Journal of Financial Economics*, 2019, 134 (3), 501–524.
- Koijen, Ralph S. J. and Motohiro Yogo**, “A Demand System Approach to Asset Pricing,” *Journal of Political Economy*, 2019, 127 (4), 1475–1515.

- Koijen, Ralph S. J., François Koulischer, Benoît Nguyen, and Motohiro Yogo**, “Inspecting the Mechanism of Quantitative Easing in the Euro Area,” *Journal of Financial Economics*, 2021, *140* (1), 1–20.
- Koijen, Ralph S. J., Robert J. Richmond, and Motohiro Yogo**, “Which Investors Matter for Equity Valuations and Expected Returns?,” *Review of Economic Studies*, 2024, *91* (4), 2387–2424.
- Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh**, “Shrinking the Cross-Section,” *Journal of Financial Economics*, 2020, *135* (2), 271–292.
- Lee, Charles M. C., Terrence Tianshuo Shi, Stephen Teng Sun, and Ran Zhang**, “Production Complementarity and Information Transmission across Industries,” *Journal of Financial Economics*, 2024, *155* (103812), 1–22.
- Lettau, Martin and Markus Pelger**, “Factors That Fit the Time Series and Cross-Section of Stock Returns,” *Review of Financial Studies*, 2020, *33* (5), 2274–2325.
- Madhavan, Ananth, Aleksander Sobczyk, and Andrew Ang**, “What Happens With More Funds Than Stocks?,” *Journal of Investment Management*, 2021, *19* (2), 4–28.
- McInnes, Leland, John Healy, and James Melville**, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” 2020. Working paper.
- Michel, Paul, Omer Levy, and Graham Neubig**, “Are Sixteen Heads Really Better than One?,” in H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, Vol. 32 2019.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean**, “Distributed Representations of Words and Phrases and Their Compositionality,” in C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems*, Vol. 26 2013.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean**, “Efficient Estimation of Word Representations in Vector Space,” in “International Conference on Learning Representations” 2013.
- Nagel, Stefan**, *Machine Learning in Asset Pricing* Princeton Lectures in Finance, Princeton, NJ: Princeton University Press, 2021.

- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning**, “GloVe: Global Vectors for Word Representation,” in Alessandro Moschitti, Bo Pang, and Walter Daelemans, eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics Doha, Qatar 2014, pp. 1532–1543.
- Prince, Simon J. D.**, *Understanding Deep Learning*, Cambridge, MA: MIT Press, 2023.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever**, “Improving Language Understanding by Generative Pre-Training,” 2018. Working paper.
- Reimers, Nils and Iryna Gurevych**, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” in Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics Hong Kong 2019, pp. 3982–3992.
- Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer**, “High-Resolution Image Synthesis with Latent Diffusion Models,” in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition” 2022, pp. 10684–10695.
- Sarkar, Suproteem**, “Economic Representations,” 2025. Working paper.
- Sarkar, Suproteem and Keyon Vafa**, “Lookahead Bias in Pretrained Language Models,” 2024. Working paper.
- S&P Global**, *Compustat Fundamentals* 2024.
- Stein, Jeremy C.**, “Sophisticated Investors and Market Efficiency,” *Journal of Finance*, 2009, 64 (4), 1517–1548.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin**, “Attention Is All You Need,” in I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, Vol. 30 2017.

Appendix

A. Proofs for the Model of Portfolio Holdings as Embeddings Data

We derive all results for the asset pricing model in Section 2.

A.1. Equilibrium Asset Prices

Since we normalized the shares outstanding of each asset to 1, market clearing for asset a is

$$P_a = \sum_{i=1}^I H_{ia}. \quad (\text{A1})$$

We solve for the equilibrium asset prices by linearizing the market clearing equation (A1) and substituting asset demand (4) and (5). We approximate investor i 's log dollar holding of asset a around $\bar{h}_{ia} = \ln(V_i w_a)$, where V_i is investor i 's wealth and w_a are portfolio weights (e.g., market weights) that sum to one (i.e., $\sum_{a=1}^A w_a = 1$). We define the corresponding log market capitalization that satisfies market clearing as $\bar{p}_a = \ln\left(\sum_{i=1}^I V_i w_a\right)$. A first-order approximation of the logarithm of the market clearing equation (4) is

$$\exp(\bar{p}_a) (1 + p_a - \bar{p}_a) = \sum_{i=1}^I \exp(\bar{h}_{ia}) (1 + h_{ia} - \bar{h}_{ia}). \quad (\text{A2})$$

Let $S_i = \frac{V_i}{\sum_{j=1}^I V_j}$ be investor i 's wealth as a share of total wealth. We substitute asset demand (4) and (5) in equation (A2) and solve for the equilibrium asset prices:

$$p_a = \left(\frac{\mathbf{\Lambda}_S}{\zeta_S}\right)' \mathbf{x}_a + \alpha + \frac{\xi_{Sa}}{\zeta_S}, \quad (\text{A3})$$

where $\zeta_S = \sum_{i=1}^I S_i \zeta_i$, $\mathbf{\Lambda}_S = \sum_{i=1}^I S_i \mathbf{\Lambda}_i$, $\xi_{Sa} = \sum_{i=1}^I S_i \xi_{ia}$, and

$$\alpha = \frac{1}{\zeta_S} \left(\sum_{i=1}^I S_i (\kappa_i - v_i) + \ln \left(\sum_{i=1}^I V_i \right) \right). \quad (\text{A4})$$

A.2. Reduced-Form Demand

We substitute the equilibrium asset price (A3) in asset demand (4) to obtain reduced-form demand (6), where

$$\boldsymbol{\lambda}_i = \boldsymbol{\Lambda}_i - \frac{\zeta_i}{\zeta_S} \boldsymbol{\Lambda}_S, \quad (\text{A5})$$

$$\delta_i = \kappa_i + (1 - \zeta_i) \alpha, \quad (\text{A6})$$

$$\delta_a = \frac{\boldsymbol{\Lambda}'_S \mathbf{x}_a + \xi_{Sa}}{\zeta_S}, \quad (\text{A7})$$

$$\epsilon_{ia} = \xi_{ia} - \frac{\zeta_i}{\zeta_S} \xi_{Sa}. \quad (\text{A8})$$

The investor embedding (A5) is investor i 's semi-elasticity of the portfolio holding to the asset embedding, relative to the average semi-elasticity of the other investors. This relative semi-elasticity includes an adjustment $\frac{\zeta_i}{\zeta_S}$ for the relative price elasticity. Heterogeneity in either $\boldsymbol{\Lambda}_i$ or ζ_i implies heterogeneity in $\boldsymbol{\lambda}_i$ across investors.

The residual (A8) has a factor structure through the term $\frac{\zeta_i}{\zeta_S} \xi_{Sa}$, which represents investors trading against the idiosyncratic demand shocks of other investors. There are two assumptions under which we have identification of the asset embeddings in reduced-form demand (6). First, reduced-form demand is a pure factor model when the price elasticities ζ_i are constant across investors. In this case, the investor fixed effect δ_a absorbs the size-weighted idiosyncratic demand shifter ξ_{Sa} . Second, we have identification if the asset embeddings are uncorrelated with the size-weighted idiosyncratic demand shifters. A special case of this assumption is the limit of infinitely many atomistic investors such that the size-weighted idiosyncratic demand shifter converges to zero.

We emphasize that this holdings-based approach allows us to estimate $\boldsymbol{\lambda}_i$, and $\mathbf{x}_a$, which are relative concepts (with cross-sectional mean equal to 0), not absolute ones. For example, in the simple case of uniform ζ_i , we will measure $\boldsymbol{\lambda}_i^e = \boldsymbol{\Lambda}_i - \boldsymbol{\Lambda}_E$, which is the relative taste of investor i for characteristics (how much more he likes a characteristics, compared to the average investor with taste $\boldsymbol{\Lambda}_E = \frac{1}{I} \sum_{i=1}^I \boldsymbol{\Lambda}_i$), not the absolute taste $\boldsymbol{\Lambda}_i$. Likewise, $\mathbf{x}_{ak}$ is not the “absolute” determinant of growth rate but rather the relative one: how much the firm is expected to grow, compared to the rest of the market.

Prices are what gives us the key information about the average taste $\boldsymbol{\Lambda}_S$ for characteristics. To get it (up to a multiplicative constant), we run $p_a = \gamma x_a + \dots$, as in (14), and get $\gamma = \frac{\boldsymbol{\Lambda}_S}{\zeta_S}$ (see (A3)). Hence, the philosophy in the present approach is to do as much as possible without estimating prices elasticities by instrumental variables, something always a bit delicate.

A.3. Asset Embeddings through Returns, Volume, and Portfolio Rebalancing

If an asset does not pay dividends, its log return from time $t - 1$ to t is $r_{at} = \Delta p_{at}$. Furthermore, we assume that the asset embeddings are stable from time $t - 1$ to t , so that $\mathbf{x}_{at} \approx \mathbf{x}_{a,t-1}$. We then write equation (A3) in first differences as

$$r_{at} = \left(\frac{\Delta \mathbf{\Lambda}_{St}}{\zeta_S} \right)' \mathbf{x}_{a,t-1} + \frac{\Delta \xi_{Sat}}{\zeta_S}. \quad (\text{A9})$$

We could use this equation to estimate the asset embeddings as the factor loadings in a factor model of returns. We could use returns at any frequency (e.g., daily, monthly, or quarterly) or even volume to estimate the asset embeddings, as long as they are stable.

If the asset embeddings are stable from time $t - 1$ to t , we write reduced-form demand (6) in first differences as

$$\Delta h_{iat} = \Delta \boldsymbol{\lambda}'_{it} \mathbf{x}_{a,t-1} + \Delta \delta_{it} + \Delta \delta_{at} + \Delta \epsilon_{iat}. \quad (\text{A10})$$

Equations (A9) and (A10) imply that we could estimate asset embeddings based on portfolio rebalancing, defined as $\Delta q_{iat} = \Delta h_{iat} - r_{at}$. Portfolio rebalancing could be more informative than portfolio holdings in the presence of inertia, capital gains taxes, or portfolio constraints.

B. Estimating the Recommender System by Alternating Least Squares

We describe the estimation of the RS-L model by alternating least squares.¹³ We generalize our description to supervised estimation of the asset embeddings. The estimation sample consists of only the positive positions. For log dollar holdings, we denote the sample size as $N_h = \sum_{i=1}^I |\mathcal{N}_i|$ and the variance as $\sigma_h^2 = \text{Var}(h_{ia})$. For the target variable, we denote the sample size as $N_y = MA$ and the variance as $\sigma_y^2 = \text{Var}(y_{am})$. We define $\phi_h = \frac{1-\phi}{N_h \sigma_h^2}$ and $\phi_y = \frac{\phi}{N_y \sigma_y^2}$, where $\phi \in [0, 1)$ is the relative weight on the target task.

We estimate the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\lambda}'_i, \mathbf{x}'_a, \delta_i, \delta_a, \boldsymbol{\beta}'_m, \delta_m)'$ through the objective function:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \phi_h \sum_{i=1}^I \sum_{a \in \mathcal{N}_i} (h_{ia} - \boldsymbol{\lambda}'_i \mathbf{x}_a - \delta_i - \delta_a)^2 &+ \phi_y \sum_{m=1}^M \sum_{a=1}^A (y_{am} - \boldsymbol{\beta}'_m \mathbf{x}_a - \delta_m)^2 \\ &+ \frac{\eta_\lambda}{IK} \sum_{i=1}^I \boldsymbol{\lambda}'_i \boldsymbol{\lambda}_i + \frac{\eta_x}{AK} \sum_{a=1}^A \mathbf{x}'_a \mathbf{x}_a + \eta_\beta \sum_{m=1}^M \boldsymbol{\beta}'_m \boldsymbol{\beta}_m. \end{aligned} \quad (\text{B1})$$

¹³We have also considered the Adam algorithm in the JAX machine learning library as an alternative estimation methodology.

The first term is the mean squared error for log dollar holdings. The second term is the mean squared error for the target variable. The remaining terms regularize the investor embeddings, asset embeddings, and the parameters in the target task. We could generalize the objective function to also regularize the fixed effects $(\delta_i, \delta_a, \delta_m)$. We estimate the regularization parameters $(\phi, \eta_\lambda, \eta_x, \eta_\beta)$ by ten-fold cross-validation.

Let $\boldsymbol{\lambda}_i^\delta = (\boldsymbol{\lambda}'_i, \delta_i)'$, $\bar{\mathbf{x}}_a = (\mathbf{x}'_a, 1)'$, $\bar{\boldsymbol{\lambda}}_i = (\boldsymbol{\lambda}'_i, 1)'$, $\mathbf{x}_a^\delta = (\mathbf{x}'_a, \delta_a)'$, $\bar{\boldsymbol{\beta}}_m = (\boldsymbol{\beta}'_m, 0)'$, and $\boldsymbol{\beta}_m^\delta = (\boldsymbol{\beta}'_m, \delta_m)'$. We rewrite the objective function (B1) as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \phi_h \sum_{i=1}^I \sum_{a \in \mathcal{N}_i} (h_{ia} - \delta_a - \boldsymbol{\lambda}_i^{\delta'} \bar{\mathbf{x}}_a)^2 &+ \phi_y \sum_{m=1}^M \sum_{a=1}^A (y_{am} - \boldsymbol{\beta}_m^{\delta'} \bar{\mathbf{x}}_a)^2 \\ &+ \frac{\eta_\lambda}{IK} \sum_{i=1}^I \boldsymbol{\lambda}'_i \boldsymbol{\lambda}_i + \frac{\eta_x}{AK} \sum_{a=1}^A \mathbf{x}'_a \mathbf{x}_a + \eta_\beta \sum_{m=1}^M \boldsymbol{\beta}'_m \boldsymbol{\beta}_m. \end{aligned} \quad (\text{B2})$$

Let $\mathbb{1}(a \in \mathcal{N}_i)$ be an indicator function that is equal to one if $a \in \mathcal{N}_i$. We define the vectors $\bar{\mathbf{h}}_i = [h_{ia} - \delta_a]_{a \in \mathcal{N}_i}$, $\bar{\mathbf{h}}_a = [\mathbb{1}(a \in \mathcal{N}_i)(h_{ia} - \delta_i)]_{i=1}^I$, $\bar{\mathbf{y}}_a = [y_{am} - \delta_m]_{m=1}^M$, and $\mathbf{y}_m = [y_{am}]_{a=1}^A$. We define the matrices $\bar{\mathbf{x}} = [\bar{\mathbf{x}}'_a]_{a=1}^A$, $\bar{\mathbf{x}}_i = [\bar{\mathbf{x}}'_a]_{a \in \mathcal{N}_i}$, $\bar{\boldsymbol{\lambda}} = [\bar{\boldsymbol{\lambda}}'_i]_{i=1}^I$, and $\bar{\boldsymbol{\beta}} = [\bar{\boldsymbol{\beta}}'_m]_{m=1}^M$. Let $\mathbf{D}(\eta)$ be a diagonal matrix with the elements η , except that the last diagonal element is zero.

We initialize the parameters at the estimated values for the RS-L-0 model. We then iterate on the following steps until convergence.

1. The first-order condition for $\boldsymbol{\lambda}_i^\delta$ is

$$-\phi_h \sum_{a \in \mathcal{N}_i} (h_{ia} - \delta_a - \boldsymbol{\lambda}_i^{\delta'} \bar{\mathbf{x}}_a) \bar{\mathbf{x}}_a + \mathbf{D}\left(\frac{\eta_\lambda}{IK}\right) \boldsymbol{\lambda}_i^\delta = \mathbf{0}, \quad (\text{B3})$$

which implies that

$$\boldsymbol{\lambda}_i^\delta = \left(\phi_h \bar{\mathbf{x}}'_i \bar{\mathbf{x}}_i + \mathbf{D}\left(\frac{\eta_\lambda}{IK}\right) \right)^{-1} \phi_h \bar{\mathbf{x}}'_i \bar{\mathbf{h}}_i. \quad (\text{B4})$$

2. The first-order condition for $\boldsymbol{\beta}_m^\delta$ is

$$-\phi_y \sum_{a=1}^A (y_{am} - \boldsymbol{\beta}_m^{\delta'} \bar{\mathbf{x}}_a) \bar{\mathbf{x}}_a + \mathbf{D}(\eta_\beta) \boldsymbol{\beta}_m^\delta = \mathbf{0}, \quad (\text{B5})$$

which implies that

$$\boldsymbol{\beta}_m^\delta = (\phi_y \bar{\mathbf{x}}' \bar{\mathbf{x}} + \mathbf{D}(\eta_\beta))^{-1} \phi_y \bar{\mathbf{x}}' \mathbf{y}_m. \quad (\text{B6})$$

3. The first-order condition for $\mathbf{x}_a^\delta$ is

$$-\phi_h \sum_{i=1}^I \mathbb{1}(a \in \mathcal{N}_i) \left(h_{ia} - \delta_i - \bar{\mathbf{X}}_i' \mathbf{x}_a^\delta \right) \bar{\mathbf{X}}_i - \phi_y \sum_{m=1}^M \left(y_{am} - \delta_m - \bar{\boldsymbol{\beta}}_m' \mathbf{x}_a^\delta \right) \bar{\boldsymbol{\beta}}_m + \mathbf{D} \left(\frac{\eta_x}{AK} \right) \mathbf{x}_a^\delta = \mathbf{0}, \quad (\text{B7})$$

which implies that

$$\mathbf{x}_a^\delta = \left(\phi_h \bar{\mathbf{X}}' \bar{\mathbf{X}} + \phi_y \bar{\boldsymbol{\beta}}' \bar{\boldsymbol{\beta}} + \mathbf{D} \left(\frac{\eta_x}{AK} \right) \right)^{-1} \left(\phi_h \bar{\mathbf{X}}' \bar{\mathbf{h}}_a + \phi_y \bar{\boldsymbol{\beta}}' \bar{\mathbf{y}}_a \right). \quad (\text{B8})$$

4. The embeddings are identified only up to scaling and rotation. As an optional step, we scale and rotate the asset embeddings to satisfy

$$\frac{1}{A} \mathbf{x}' \mathbf{x} = \mathbf{I}_K. \quad (\text{B9})$$

Thus, $\frac{1}{A} \sum_{a=1}^A x_{ak}^2 = 1$ for each component k , so that the scaling does not depend on the number of assets.

B.1. Out-of-Sample Testing

In supervised estimation, we are often interested in how well the asset embeddings perform out of sample. In this case, we split the sample of the target variable into training and testing data. We then use only the training data to estimate the asset embeddings.

C. Extensions of the AI Methods

C.1. Recommender Systems Using Historical Data

We estimate the recommender system on each cross section. We could connect the asset embeddings across quarters by adding another regularization term $\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$ to the objective function (7). This term stabilizes the unidentified rotation across quarters and limits the time variation in the asset embeddings. The intuition is that the asset embeddings in the previous quarter serve as a prior for the current asset embeddings. We could of course add more lags and parameters to enrich the time-series dynamics of the asset embeddings. An advantage of using historical data is that we can sharpen the estimates of asset and investor embeddings. Moreover, by estimating asset embeddings that can be compared across periods, we can use them for return predictability.

C.2. Estimating Asset Embeddings by GloVe

GloVe takes the co-occurrence matrix of words across documents as an input, based on how frequently pairs of words appear in a context window. GloVe estimates the embeddings by minimizing a weighted least squares objective, which captures the difference between the dot product of the embeddings and the logarithm of the frequency of co-occurrences. We could adopt this method by constructing a co-occurrence matrix of assets across portfolios, based on how frequently pairs of assets appear in a context window.

C.3. Transformer Models Using Historical Data

Unlike word embeddings in NLP, asset and investor embeddings change at a higher frequency. Some characteristics like industry classification are highly persistent over time, while other characteristics like profitability and risk exposure change at a higher frequency. In the baseline specification, we train the BERT model on each cross section, which is simple but potentially inefficient. Therefore, we consider using historical data to estimate the asset embeddings. When using historical data on holdings beyond the current quarter, we restrict attention to the list of stocks in the current quarter as the vocabulary.

To explore this issue, we consider a generalized training procedure that proceeds in two steps. First, we use a longer sample up to five years to train the BERT model. Second, we fine-tune the model on one to four quarters of the most recent data. The first step captures the persistent component of asset embeddings, and the second step captures the higher frequency variation. For the managed portfolio benchmark, we found a performance gain of up to 5% in the out-of-sample R^2 . Because of the limited performance gain, we present the baseline specification in the paper to avoid the additional complexity.

There are other approaches to introduce the historical data, which we leave for future research. First, we currently mask 15% of the sample to construct the training data. We could use the historical data but decrease the masking share to put less weight on the older data. Second, we could examine the estimated parameters in the BERT model to find out which ones vary significantly over time. We could then use the historical data but allow only a subset of the parameters to vary over time. This procedure is similar to the low rank approximation (LoRa) approach to fine-tuning LLMs.

C.4. Integrated Model of Asset and Investor Embeddings

We separately estimate the PS-BERT model for the investor embeddings and the OS-BERT model for the asset embeddings. We could combine them into a single model and also incorporate log dollar holdings in the objective function, instead of the ranked positions as

we currently do. We start with the input embeddings for investors $\mathbf{\lambda}_i$ and assets $\mathbf{x}_a$. For each asset a , we estimate the contextualized investor embedding $\mathbf{\lambda}_i^a$ as a weighted average of the input embeddings for the other investors holding asset a . For each investor i , we estimate the contextualized asset embedding $\mathbf{x}_a^i$ as a weighted average of the input embeddings for the other assets in investor i 's portfolio. We then train the BERT model to minimize the objective function:

$$\sum_{i=1}^I \sum_{a=1}^A (h_{ia} - \mathbf{\lambda}_i^{a'} \mathbf{x}_a^i - \delta_i - \delta_a)^2, \quad (\text{C1})$$

subject to the same regularization methods as the traditional BERT model.

D. Data Construction

D.1. Stock Market and Portfolio Holdings Data

We provide a detailed description of how we construct the stock market and portfolio holdings data.

D.1.1. FactSet Ownership Data

We construct the quarterly stock holdings of mutual funds, ETFs, closed-end funds, variable annuity funds, and hedge funds using the FactSet Ownership Data (FactSet, 2024). In the 13F Ownership Data, FactSet provides the quarter-end reported dollar holding (adj_mv) by 13F filer, CUSIP, and report date. We aggregate the dollar holding up to the rollup entity level. We keep only hedge funds, based on the entity sub-type in the own_ent_institutions file.

In the Fund Ownership Data, FactSet provides the reported dollar holding (adj_mv) by fund identifier, CUSIP, and report date. Funds often report at a frequency higher than quarterly. We construct the quarter-end dollar holding by keeping the last report in each quarter when sorted by report date, filing date, transfer date, and form type. We keep observations where the report date is within five days of the quarter-end date. Based on the fund type, we keep only mutual funds, ETFs, closed-end funds, and variable annuity funds.

We construct market equity as the product of unadjusted price and unadjusted shares outstanding. We merge the dollar holdings and market equity by date and CUSIP and keep observations where both variables are positive.

D.1.2. Firm Characteristics and Stock Returns

We use the data on firm characteristics and stock returns from Jensen et al. (2023), which are based on the CRSP US Stock Database (Center for Research in Security Prices, 2024) and Compustat Fundamentals (S&P Global, 2024). We use book equity (be), market equity (me), the market beta (beta_60m), book-to-market equity (be_me), asset growth (at_gr1), profitability (gp_at), momentum (ret_12_1), and the dividend-asset ratio (div_at).

The data on firm characteristics and stock returns are at the security (CRSP permno) level. We aggregate them to the firm (CRSP permco) level, using market equity weights. Jensen et al. (2023) group stocks into five size categories. We keep firms for which at least one of its stocks is a small cap or above (i.e., above the 20th percentile for market equity). We drop any firms with missing characteristics or non-positive book equity. We winsorize log book-to-market equity, asset growth, profitability, and momentum at the 1st and 99th percentiles in each cross section. We winsorize the dividend-asset ratio at the 99th percentile in each cross section.

D.1.3. Merging and Cleaning

We first merge the FactSet stock holdings data with the CRSP-Compustat link file by date and historical CUSIP. We use the historical CUSIP from the FactSet sym_cusip_hist file. We then merge the data on firm characteristics and stock returns from Jensen et al. (2023) by date and CRSP permco.

Our sample starts in 2005.Q1, which is dictated by a limited coverage of funds in the FactSet Fund Ownership Data before that date. We drop investors with a highly concentrated portfolio, defined as a single holding that exceeds 75% of the portfolio. In each cross section, we keep investors who hold at least 20 stocks and stocks that are held by at least 20 investors, which we impose by iterating on these two criteria until convergence.

We winsorize log dollar holdings. In each cross section, we first remove investor fixed effects by subtracting the median by investor. We then remove stock fixed effects by removing the median by stock. We then winsorize these centered log dollar holdings at the 2.5th and 97.5th percentiles. Finally, we recenter the winsorized log dollar holdings by removing the investor and stock fixed effects, using the mean instead of the median.

D.2. Text-Based Embeddings

We use the text-based asset embeddings from Cohere and OpenAI, based on the most recent publicly available models as of March 2024. We use the list of CRSP company names to

download the embeddings as of 2022.Q4. Following the recommendations of Cohere and OpenAI, we use cosine similarity to measure firm similarity.

We use the embed-english-v3.0 model from Cohere. Cohere provides task-specific embeddings for search query, search document, classification, and clustering. We use the clustering embeddings since our application is to identify similar firms. The embeddings have 1,024 dimensions. For the relative valuation benchmark, we reduce the embeddings to four and ten dimensions by Uniform Manifold Approximation and Projection (McInnes et al., 2020).

We use the text-embedding-3-large model from OpenAI. OpenAI allows us to download their embeddings at any desired dimension up to 3,072. Thus, we download the embeddings at four and ten dimensions for the relative valuation benchmark.