TUM School of Management
Technische Universität München

Leveraging the Latest Advances in NLP and AI for Asset Pricing

Christian Breitung

Vollständiger Abdruck der von der TUM School of Management der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Martin Meißner

Prüfende der Dissertation:

1. Prof. Dr. Sebastian Müller
2. Prof. Dr. Reiner Braun

Die Dissertation wurde am 22.04.2025 bei der Technischen Universität München eingereicht und durch die TUM School of Management am 15.06.2025 angenommen.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Sebastian Müller, for his continuous support, invaluable guidance, and thoughtful feedback throughout this dissertation. His encouragement to independently pursue new research ideas, combined with his openness to regular and constructive discussions, has been instrumental in shaping this work. I am especially grateful for the academic freedom he provided, which allowed me to develop my research interests in a supportive and intellectually stimulating environment. I also sincerely appreciate his support for my research visits to Boston College and the University of Southern California. These opportunities broadened my academic horizons and enabled valuable exchanges with leading scholars, enriching my international experience.

My sincere thanks also go to my examination committee, particularly to my second reviewer, Prof. Dr. Reiner Braun, for his constructive comments and critical reflections, which helped strengthen this thesis. I am equally grateful to the chair of the committee, Prof. Dr. Martin Meißner, for his support and engagement in the evaluation process.

I extend my appreciation to my co-authors for their collaborative spirit, insightful contributions, and the many enriching discussions we shared, from initial idea generation to resolving complex technical challenges. I also thank my colleagues for their feedback during internal seminars and informal exchanges, which contributed significantly to the refinement of this work. I would also like to express my sincere gratitude to Prof. Dr. Vitor Gonçalves de Azevedo for serving as the mentor of my dissertation.

I am profoundly thankful to my parents, whose unwavering support, encouragement, and early guidance laid the foundation for my academic path. Your values, patience, and belief in me have shaped both my professional and personal development. I also want to express my heartfelt thanks to my brother. You have been a role model and source of inspiration from an early age.

Finally, I would like to thank my close friends for their support, understanding, and encouragement throughout this journey. Your belief in me has been a constant source of motivation.

Abstract

Efficient stock markets rely on investors' capacity to rapidly process and accurately distinguish between value-relevant and irrelevant information. However, with the exponential growth in quantitative and qualitative data, traditional methods of information processing face significant challenges.

This thesis applies Artificial Intelligence (AI) and Natural Language Processing (NLP) to examine how effectively investors incorporate new information into global asset prices. First, it demonstrates the potential of machine learning models, particularly random forests, to systematically identify predictable patterns in the cross-section of global stock returns. Second, it explores how text embeddings can be used to detect previously unpriced information in corporate disclosures, thereby shedding light on patterns of investor inattention across global markets. Third, it employs Large Language Models (LLMs) to analyze macroeconomic reports, revealing that investors often fail to immediately and fully incorporate the information contained in these publications. Fourth, the thesis constructs a global business network based on the textual analysis of firm disclosures. This network can improve peer identification, industry classification, and provides a more detailed view of economic links across the globe. Finally, it introduces a novel approach that combines embedding models with topic modeling to generate interpretable textual representations. The advantages of these representations are showcased by deriving more accurate and interpretable estimations for factor betas.

Overall, the findings of this thesis underscore how recent advances in AI and NLP can support investors in processing information more effectively and pricing securities more efficiently on a global scale. By integrating modern computational tools into financial analysis, this thesis offers both methodological innovations and new empirical insights into the dynamics of information incorporation in capital markets.

Contents

List of Abbreviations

ACE Aggregated Cluster Embeddings

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

CAPM Capital Asset Pricing Model

EMH Efficient Market Hypothesis

FED Federal Reserve System

GenAI Generative Artificial Intelligence

ISI Industry-Specific Impact

LDA Latent Dirichlet Allocation

LLM Large Language Model

LLMs Large Language Models

LSEG London Stock Exchange Group

ML Machine Learning

NID New Information Detection

NLP Natural Language Processing

OECD Organisation for Economic Co-operation and Development

SEC Securities and Exchange Commission

SR Sharpe ratio

Publication and Submission Record

This publication-based thesis comprises five papers. The first and fourth are published, while the others are working papers. I am the the corresponding and leading author for all five paper.

Breitung, C. (2023)
Automated Stock Picking using Random Forests
Journal of Empirical Finance, Volume 72, Pages 532-556
https://doi.org/10.1016/j.jempfin.2023.05.001
VHB Journal Ranking 2025: B

Breitung, C. & Müller S. (2022)
New Information Detection using Language Models
Working Paper

Breitung, C., Kruthof G. & Müller S. (2025)
Macroeconomic Inattention
Working Paper

Breitung, C. & Müller S. (2025)
Global Business Networks
Journal of Financial Economics, April 2025, Article 104007
https://doi.org/10.1016/j.jfineco.2025.104007
VHB Journal Ranking 2025: A+

Breitung, C. (2025)
Aggregated Cluster Embeddings
Working Paper

1 Introduction

According to the Efficient Market Hypothesis (EMH), a market is considered efficient if asset prices fully and instantaneously reflect all publicly available information (Fama, 1970). This foundational assumption implies that market participants are not only rational but also capable of processing new information swiftly and accurately translating its implications into revised asset valuations.

In practice, however, the range of potentially value-relevant data is vast and heterogeneous. While traditional models emphasize quantitative sources such as historical stock prices, balance sheet ratios, and earnings metrics, qualitative information, such as forward-looking statements in corporate disclosures, the tone of earnings calls, or the content of risk factor narratives, may also affect the valuation of companies. Moreover, unconventional signals from news media or social media platforms increasingly shape investor sentiment and short-term return dynamics.

Next to firm-specific information on the focal firm, developments among economically connected firms, such as suppliers, customers, or competitors, can also affect the valuation of the focal company. For instance, a supply chain disruption at a key supplier may propagate downstream effects that impair production, increase costs, or delay revenues for dependent firms.

Similarly, macroeconomic indicators, including interest rate projections, inflation expectations, employment data, and consumer sentiment indices, further complicate the information landscape. These variables directly affect firm valuations by altering assumptions about cash flow growth. Furthermore, they can also indirectly affect discount rates through monetary policy channels.

The central challenge, therefore, lies in distinguishing meaningful signals from the large amount of publicly available information. Not all data are informative, timely, or relevant. Reporting biases and market narratives may make many signals redundant, outdated, or distorted. Manual analysis becomes increasingly infeasible as the complexity and volume of available data continue to rise. In such an environment, the limits of investor attention and processing capacity may

give rise to systematic underreaction, informational inefficiencies, and persistent mispricings, posing a fundamental challenge to the strong-form interpretation of the EMH.

Traditionally, the analysis of market efficiency has relied heavily on quantitative data, such as historical stock prices and firm fundamentals. Within this framework, researchers have documented a vast and growing array of return anomalies, referred to as the "zoo of new factors" (Cochrane, 2011), that cannot be reconciled with classical asset pricing models like the Capital Asset Pricing Model (CAPM) or the Fama-French three- or five-factor models.

Motivated by the many anomalies, researchers have increasingly turned to Machine Learning (ML) to uncover high-dimensional patterns in asset returns. ML models are particularly well-suited for this task because they can flexibly capture high-dimensional relationships and non-linear effects without relying on strong parametric assumptions. A seminal contribution in this area is provided by Gu et al. (2020), who apply a variety of ML algorithms, such as random forests and neural networks, to a large set of return predictors drawn from the academic literature. Their findings reveal substantial out-of-sample predictive power in the cross-section of U.S. stock returns, suggesting that financial markets may exhibit a greater degree of inefficiency than previously assumed.

While these results are promising, they also raise concerns about economic relevance and implementability. Avramov et al. (2023) argue that the predictive power of ML models is primarily driven by stocks that are more volatile, less liquid, and costlier to trade. This casts doubt on the feasibility of exploiting these signals in practice. Nonetheless, their study finds that even long-only portfolios, which are more straightforward to implement, can deliver abnormal returns when constructed using ML-generated signals. This underscores the potential for ML-based strategies to offer economically meaningful insights despite the frictions present in real-world markets. By allowing researchers to systematically analyze large sets of anomalies and explore their interactions without imposing restrictive assumptions, ML models open new ways to understand market inefficiencies and improve return forecasts.

In addition to the analysis of quantitative data, researchers have increasingly turned their attention to the role of textual information in financial markets, aiming to assess whether such qualitative data is efficiently reflected in asset prices. This growing interest is driven by the recognition that a large share of the value-relevant information in financial markets is conveyed through corporate disclosures, earnings calls, news articles, analyst reports, and online forums, which require textual analysis methods.

One of the earliest influential contributions to the use of textual data in finance is provided by Antweiler and Frank (2004), who examine whether stock-related messages posted on Yahoo Finance and Raging Bull forums influence market activity. Building on this line of research, Tetlock (2007) apply a psychological dictionary to assess the tone of media articles and explore its relationship with market returns, while Tetlock et al. (2008) extend this approach to predict firms' accounting earnings and stock returns using similar sentiment metrics. However, these early methods rely on general-purpose dictionaries developed outside the financial context, which may misinterpret domain-specific language. To address this, Loughran and McDonald (2011) develop finance-specific dictionaries containing refined sets of positive, negative, and neutral terms better suited for analyzing financial text.

More recent research, such as Frankel et al. (2022), has moved beyond static word lists by employing machine learning techniques to extract meaningful one- and two-word phrases from 10-K filings. Their study demonstrates that this machine learning-based approach significantly outperforms the traditional dictionary method of Loughran and McDonald (2011), especially in recent years. This improvement may be linked to linguistic shifts documented by Cao et al. (2023), who find that firms increasingly tailor their disclosure language to avoid negatively connoted words from widely used dictionaries, thereby reducing the effectiveness of static lexicons for sentiment analysis.

Next to sentiment analysis, textual analysis can also be used to assess the value-relevance of individual news articles. For instance, by employing text regression techniques, Boudoukh et al. (2019) demonstrate that price movements are more pronounced when associated with news that contain more foundamental information. Furthermore, Fedyk (2024) analyze the effect of news positioning on financial markets by analyzing Bloomberg terminal data. The study found that articles featured on the front page experience significantly higher trading volumes and larger absolute excess returns shortly after publication compared to equally important articles not given front-page prominence.

Finance researchers also use similarity methods to investigate textual data. Here, the most applied method is the bag-of-words approach. Text documents are represented as vectors, where each dimension in the vector represents an individual word. Hoberg and Phillips (2010, 2016) use bag-of-words to identify firms who operate in similar product markets and Andreou et al. (2020) use it to identify firms with similar market orientation. Brown and Tucker (2011) applies bag-of-words when investigating changes in the MD&A section of 10-K filings, and

Cohen et al. (2020) use this method to test whether investors timely price information that is contained in corporate disclosures.

Topic modeling is another methodology that researchers frequently apply to finance texts. For instance, Campbell et al. (2014) use word counts to assess the importance of disclosed risk factors in U.S. annual reports. In contrast, Lopez-Lira (2023) uses Latent Dirichlet Allocation (LDA) to identify risk clusters in 10-K filings, constructs text-based factors, and shows that these factors perform at least as well as classic asset pricing models. Bybee et al. (2024) develop a narrative asset pricing model by extracting systematic risk factors from news text using topic modeling. Their approach integrates topic modeling with latent factor analysis to identify interpretable risk factors that achieve higher out-of-sample Sharpe ratio (SR) and smaller pricing errors than traditional models.

Using word lists to quantify textual data remains a popular approach among researchers, largely due to its low computational demands and the high degree of interpretability. For sentiment analysis, these dictionaries allow for transparent mappings of textual content into predefined categories, such as positive or negative, which facilitates intuitive analyses and replicability. However, despite these advantages, dictionaries are subject to notable limitations. A dictionary approach is not able to account for the linguistic context in which individual words appear. For instance, while an increase in profit is clearly a positive development, an increase in interest payments is generally considered unfavorable. Dictionary-based methods, however, treat both cases similarly.

Moreover, these approaches struggle to interpret negations ("not profitable") and are generally insensitive to synonymous expressions. This limitation often leads to measurement error or an underestimation of the richness of financial language. The same challenges apply in the context of textual similarity. Two reports may convey identical information but appear dissimilar under a bag-of-words approach because they employ different vocabulary to describe the same concepts. Without understanding the broader sentence-level or paragraph-level structure, dictionary methods offer only a coarse approximation of meaning.

The rise in computational power and parallel advances in computer science have encouraged researchers in the field of NLP to explore more sophisticated methods for text analysis that overcome the limitations of traditional word-level models. An important innovation in this regard has been the development of embedding models, which allow for the representation of words, phrases, or entire documents as dense numerical vectors in a high-dimensional space.

These representations enable the modeling of semantic similarity by placing contextually related terms in close proximity to each other in vector space.

Early implementations, such as the word2vec model introduced by Mikolov et al. (2013), represent words as dense vector embeddings. Although these models do not capture contextual information contained in sentences, they nevertheless marked a critical breakthrough by mapping words with similar meanings to similar embeddings.

Recently, a novel class of embeddings has emerged within the NLP literature. This development was driven by the introduction of the transformer architecture by Vaswani et al. (2017), which enables parallel processing and significantly reduced training times. Crucially, the self-attention mechanism inherent in transformers allows models to capture the contextual meaning of words by considering their relationships within the entire sequence. This advancement has facilitated the unsupervised learning of extensive semantic representations, thereby eliminating the dependency on large, manually annotated datasets that had previously constrained the development of domain-specific models. Recognizing the potential of this new approach, major technology companies soon released transformer-based language models, including Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and RoBERTa (Liu, 2019). These encoder-based models produce context-aware text embeddings, which can be fine-tuned for domain-specific applications such as sentiment analysis or textual similarity measurement.

While BERT and its variants advanced contextual understanding through bidirectional attention mechanisms, a parallel line of research focused on decoder-only architectures optimized for generative tasks. This trajectory gained widespread attention with the release of GPT-2 (Radford et al., 2019), which showcased the potential of transformer-based language models to produce coherent and contextually relevant text over extended passages. The subsequent development of GPT-3 (Brown et al., 2020) dramatically increased model size and introduced few-shot and zero-shot learning capabilities. By conditioning the model on natural language prompts, users could generate high-quality outputs without the need for task-specific fine-tuning, thereby broadening its utility across a range of applications, including text summarization, translation, question answering, and code generation. Models such as PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), and GPT-4 (OpenAI, 2023) have continued to push the boundaries of scale, multilingual understanding, and reasoning ability. These models demonstrate improved generalization across complex tasks and domains, often matching or surpassing human performance in specific benchmark settings.

Finance researchers increasingly adopt advanced textual analysis techniques to extract economically meaningful insights from financial disclosures and news sources. These methods, powered by recent developments in NLP, allow researchers to move beyond traditional word-based representations towards more nuanced, context-aware methods. For example, Lopez-Lira and Tang (2023) employ LLMs to forecast stock price movements based on firm-specific news headlines. Leveraging the contextual understanding of LLMs, their approach evaluates both the relevance and directional impact of news on individual firms. Their results show that LLMs significantly improve the prediction of out-of-sample daily returns, outperforming standard sentiment-based models.

Similarly, Chen et al. (2022) utilize context-aware sentence embeddings to represent news headlines and demonstrate that machine learning models trained on these embeddings outperform traditional text representations in forecasting stock returns across multiple languages and countries.

Beyond return prediction, embedding models also enhance the identification of economic peers. Hoberg and Phillips (2025) propose a novel industry classification system derived from doc2vec embeddings of business descriptions, showing that their approach improves explanatory power in return comovement tests by approximately 20% relative to the TNIC dataset developed by Hoberg and Phillips (2010, 2016), which relies on bag-of-words methods.

Despite these performance advantages, modern NLP methods introduce significant challenges for empirical finance research. A key issue is the limited interpretability of pre-trained transformer models. While embedding models can effectively represent textual data, they operate as black boxes and offer no direct mapping between textual features and individual vector dimensions. This lack of transparency is particularly problematic in financial contexts, where model outputs must often be audited by analysts, investors, or regulatory bodies. The difficulty of attributing predictions to specific linguistic features limits the practical and academic value of these models when interpretability is a necessary condition.

Another critical concern is the risk of look-ahead bias. Researchers may inadvertently introduce forward-looking knowledge into past-period analyses when applying language models trained on recent financial texts to analyze historical documents. As Sarkar and Vafa (2024) emphasize, avoiding look-ahead bias in financial text applications requires careful temporal separation between training and inference. However, this challenge is compounded in practice by using commercial models that are continuously fine-tuned on the most recent corpora. As shown by Ludwig et al. (2025), even when used after their designated knowledge cutoff, such mod-

els may embed latent information that introduces bias, calling into question their validity in retrospective empirical research.

This thesis investigates how efficiently investors incorporate both quantitative and qualitative information into asset prices in global capital markets, drawing on recent advancements in AI and NLP. On the quantitative side, I examine whether the predictive performance of machine learning models documented for the U.S. generalizes to international markets. In particular, I focus on technical indicators, which, despite their widespread use among practitioners and retail traders, have received limited attention in academic finance. By systematically analyzing their predictive value across countries, the study contributes to the growing literature on alternative sources of return predictability.

On the qualitative side, I assess how efficiently investors process and price firm-specific disclosures and peer and macroeconomic information. To this end, I introduce new methodologies that quantify the economic salience of textual information while addressing the three key methodological challenges: overfitting, look-ahead bias, and lack of model interpretability. For example, I apply context-aware embeddings to detect semantically novel information in corporate reports and use explainable AI techniques to attribute predictive power to interpretable components. These innovations enable a more precise assessment of investor information processing and allow for a better understanding of why certain textual signals are priced slowly or not at all.

By integrating machine learning and advanced NLP techniques into empirical asset pricing, this thesis offers a comprehensive framework to study market efficiency in the age of big data. The findings contribute to a deeper understanding of how quantitative and qualitative signals interact in determining asset prices and shed light on the conditions under which investors fail to incorporate relevant information promptly and accurately.

2 Main Findings

2.1 Automated Stock Picking Using Random Forests

In paper 4.1, I investigate the efficiency of global stock markets by training a random forest model on technical indicators to identify outperforming stocks. Unlike conventional return prediction models, which focus on absolute returns and may be biased toward highly volatile and illiquid stocks, I introduce an alternative ranking methodology based on outperformance probabilities, which is the proportion of decision trees in a random forest that classify a stock as an outperformer.

To develop the model, I construct more than 180 technical indicators, such as Bollinger Bands, oscillators, moving averages, and momentum-based metrics, derived from global stock market data covering 68 countries from 1990 to 2018. While practitioners extensively use these indicators, they remain largely underexplored in academic finance. For each stock-month observation, I compute the next month's median return and assign outperformer or underperformer labels based on whether the return exceeds the median. These labels serve as the dependent variable in a classification framework.

I train a random forest model using the technical indicators as inputs and the binary outperformance label as the target. Outperformance probabilities are then derived as the fraction of decision trees in the ensemble that classify a stock as an outperformer. Based on these probabilities, I construct long–short portfolios that go long in high-probability stocks and short in low-probability ones. Compared to traditional regression-based approaches, the classification model avoids a systematic bias toward volatile or illiquid stocks with extreme returns, thereby producing more stable results.

The empirical findings highlight the potential of machine learning models to uncover persistent return patterns in technical indicators. Long–short portfolios sorted by outperformance probabilities achieve SR of up to 1.95 and highly significant annualized five-factor alphas as high as

21.79%. These results remain strong even in out-of-sample tests and cannot be fully explained by conventional risk factors or market frictions.

Moreover, this model-driven investment strategy is particularly profitable during market crises, with substantial outperformance observed in periods of heightened volatility. Overall, the profitability of the ranking-based strategy cannot be fully explained by traditional risk factors, transaction costs, or market inefficiencies. In particular, the strategy proves robust across regions and countries, suggesting that limits to arbitrage play only a partial role.

Using explainable AI techniques, particularly Shapley values, I assess the contribution of individual technical indicators to the model's predictions. The results reveal that trend-based indicators, such as moving averages and directional movement indices, consistently emerge as the most influential features across global markets. However, the relevance of other indicators seems to vary by region. Momentum-based signals and volume-related measures are more prominent in specific markets, reflecting heterogeneous trading behaviors and market structures across countries.

In the final step of the analysis, I evaluate the usefulness of the model's outperformance probabilities as inputs to a mean-variance portfolio optimization framework. Replacing traditional return forecasts with model-based outperformance probabilities, I construct portfolios that achieve lower out-of-sample drawdowns and volatility while preserving strong performance. These portfolios consistently exhibit higher Sharpe ratios than those built on historical return averages, demonstrating superior risk-adjusted returns. Importantly, these benefits persist even for long-only portfolios, underscoring the practical applicability of the approach. Overall, the findings suggest that machine learning-driven return estimates can enhance portfolio construction by offering more reliable, forward-looking measures of expected returns.

2.2 New Information Detection using Language Models

In paper 4.2, Sebastian Müller and I investigate whether investors promptly incorporate novel information in international annual reports into stock prices. Motivated by the increasing volume of unstructured textual data and the challenges of identifying economically relevant content across different languages and regulatory environments, we develop and implement New Information Detection (NID), a novel transformer-based framework for detecting semantically new information in corporate disclosures. In contrast to traditional word-based approaches such

as bag-of-words, NID captures contextual meaning by comparing sentence-level embeddings across consecutive reports, leveraging a fine-tuned multilingual language model.

The construction of our NID measure relies on calculating cosine similarities between sentence embeddings from two subsequent annual reports by the same firm. For each sentence in the newer report, we compute its similarity with all sentences in the prior report and record the highest score. Averaging these maximum scores yields a semantic similarity index, where a lower average indicates a higher degree of informational novelty. This embedding-based approach captures more profound contextual changes beyond superficial word substitutions and is applicable to disclosures in over 100 languages without requiring translation or language-specific preprocessing.

Applying NID to a dataset of roughly 500,000 annual reports from over 25,000 firms in more than 30 countries, we find strong evidence of investor underreaction to novel information, confirming the presence of the "lazy prices" (Cohen et al., 2020) in international stock markets. A trading strategy that takes long positions in firms with low novelty and short positions in firms with high novelty generates a statistically and economically significant six-factor alpha of 89 basis points per month globally.

To benchmark our measure against existing approaches, we compare NID to a bag-of-words novelty score in a U.S.-only setting. Both methods detect investor underreaction. A portfolio based on NID delivers higher abnormal returns (39 versus 24 basis points per month), although the difference is not statistically significant. However, in the international sample, the advantage of NID becomes more pronounced. Here, the alpha difference between NID and the bag-of-words measure is statistically significant at the 1% level. This finding highlights the benefit of transformer-based models in capturing informational novelty in less standardized and multilingual contexts.

We further analyze heterogeneity in the anomaly's magnitude across firm types and disclosure environments. The underreaction is especially pronounced among small-cap firms, which are often subject to lower investor coverage and reduced information processing. In countries with stricter regulatory frameworks, the anomaly is also amplified, consistent with the idea that more informative disclosures are more likely to be overlooked when buried in lengthy and unstructured documents. For instance, in countries where regulators possess high investigative authority, the long-short strategy yields a monthly alpha of 108 basis points, compared to just 25 in jurisdictions with weak regulatory oversight.

To ensure the robustness of our findings, we perform several placebo and matching tests. First, we examine whether data publication errors could explain the delayed market reaction by introducing lags in portfolio formation. Second, we rule out the possibility that cross-country variation in disclosure structure or NID calibration drives our results by forming portfolios within individual countries, which continue to yield significant abnormal returns. Third, we apply a matched-pairs approach, controlling for industry affiliation and firm size. We continue to observe strong and persistent return spreads in any of the specifications.

Overall, these results suggest that investors systematically underreact to semantically new content in corporate disclosures, leading to predictable return patterns that can be exploited using modern NLP tools. By introducing a context-sensitive and language-agnostic approach to quantifying textual novelty, our study contributes to the literature on investor information processing and demonstrates how advanced AI models can reveal economically meaningful inefficiencies in global capital markets.

2.3 Macroeconomic Inattention

In paper 4.3, Garvin Kruthof, Sebastian Müller, and I investigate to what extent investors efficiently process macroeconomic reports from institutions such as the Federal Reserve System (FED) and the Organisation for Economic Co-operation and Development (OECD). We introduce a novel framework that leverages LLMs to generate Industry-Specific Impact (ISI) scores for macroeconomic news. These scores quantify the expected impact of macroeconomic disclosures on different industries, providing a systematic way to assess whether investors correctly interpret such information. Applying this framework to macroeconomic reports from the FED and OECD, we implement a long–short portfolio strategy, taking long positions in industries expected to benefit from the disclosed information and short positions in industries predicted to decline. If markets were fully efficient, this strategy would not generate abnormal returns. However, empirical results reveal statistically and economically significant excess returns, with a six-factor alpha of up to 93 basis points per month, suggesting that investors systematically underreact to macroeconomic information.

To address concerns of look-ahead bias, we conduct multiple robustness tests. First, macroeconomic reports are summarized using an LLM, with all temporal markers removed, ensuring the extracted information does not reference future events. A portfolio based on these summaries generates a comparable alpha of 96 basis points per month, reinforcing that the anomaly is

not driven by look-ahead bias. Second, an analysis of macroeconomic subdimensions, including inflation, employment, and financial conditions, demonstrates that the model's weighting scheme does not optimize risk-adjusted returns. Third, a manual review of model justifications suggests that LLM-generated ISI scores rely on well-established economic principles rather than future knowledge.

Our work is motivated by prior research documenting investor underreaction to firm-level disclosures (Cohen et al., 2020). In contrast to firm-specific reports, macroeconomic reports often contain implications concerning inflation trends, consumer demand, and monetary policy that may affect industries differently. Yet, automatically extracting this information is not straightforward and requires sophisticated analysis methods. Dictionary-based methods, for example, lack the contextual awareness needed to parse nuanced economic implications, particularly across multiple dimensions and industries.

Our framework overcomes these limitations by using LLMs to extract and interpret forward-looking content in these complex documents. To validate the economic reasoning capabilities of LLMs in this context, we first evaluate their performance on a controlled dataset of 1,600 commodity news headlines. For each of the 40 commodities, we generate 40 headlines, half reflecting price increases and half decreases, using real-world linguistic patterns and manual oversight. We then ask six language models to assign industry-specific impact labels, which we benchmark against well-established economic principles. Larger models, such as GPT-4 and GPT-4o-mini, achieve an accuracy of up to 84.3%, outperforming traditional mono-directional classifiers and demonstrating their suitability for real-time macroeconomic analysis.

With this validation, we apply our LLM-based framework to full macroeconomic reports. Five days after each report's publication, we construct portfolios that go long in industries with positive ISI scores and short in those with negative ones. If markets fully and immediately priced the implications of these disclosures, the strategy should yield no excess return. However, we find robust evidence of significant alphas across various specifications. These results hold even when using masked summaries, randomly reweighted macroeconomic subdimensions, and different industry classification schemes.

We further test whether the model's results are driven by temporal leakage or look-ahead bias. For instance, we assess whether masked summaries, where all date references are omitted, still yield significant alphas. We find that the LLMs cannot reliably infer the year of the report based on these summaries, yet the investment strategy remains highly profitable, indicating that temporal information is not essential for model performance. Additionally, we examine

whether the model's subdimension weights across factors such as commodity prices, consumer sentiment, or fiscal policy systematically enhance performance. We find that these weightings are not optimized to maximize returns, which supports the claim that the model is not simply fitting noise or future-aware patterns.

Finally, we manually inspect the justifications provided by GPT-4o-mini for selected macroeconomic reports. We find that the model's reasoning aligns with standard economic theory rather than relying on future information.

2.4 Global Business Networks

In paper 4.4, Sebastian Müller and I develop a time-varying global business network that captures firm similarities across global markets. While prior research has constructed U.S.-focused networks based on textual similarities in corporate disclosures (Hoberg and Phillips, 2010, 2016), we extend this approach by introducing the first global network built from firm-level textual information, motivated by the increasing specialization and heterogeneity within industries that traditional classifications often fail to capture.

To create the network, we generate historical business descriptions for over 63,000 firms across 67 countries using annual reports obtained from the Securities and Exchange Commission (SEC) and London Stock Exchange Group (LSEG), supported by Generative Artificial Intelligence (GenAI) tools such as GPT-3. Specifically, we apply AI techniques to identify firm-specific text segments and instruct a large language model to generate concise descriptions that align with commercial business profiles.

These descriptions are then embedded using advanced embedding models from OpenAI and a complementary open-source model. Each description is transformed into a high-dimensional vector, and we define firms as economically related if the cosine similarity of their business description embeddings is above the 99th percentile of the pairwise similarity distribution. This approach generates a global business network capturing firm-level economic links both within and across borders.

Our results show that this embedding-based methodology more accurately identifies economic peers than standard industry classifications. We demonstrate that our network captures domestic and cross-border firm relationships and exhibits more substantial alignment with real-world business connections, higher return comovement, and superior predictive power in lead-lag return forecasting and M&A prediction applications. To address potential look-ahead bias from

pre-trained models like GPT-3, whose training corpus includes data up to 2021, we apply advanced masking techniques to eliminate identifiable firm information such as company names, product lines, or headquarters locations.

We further demonstrate that language models can be fine-tuned to classify economic relationships, such as distinguishing between competitors, suppliers, and customers, with over 85% accuracy based on AI-generated business descriptions. Using labeled data from FactSet Revere, we train a multiclass classifier to support more nuanced analyses of industrial networks, enabling future work on topics like supply chain resilience, input-output spillovers, and ESG diffusion.

To illustrate the network's usefulness, we apply it to two well-known tasks in the finance literature. First, we analyze the lead-lag effect in stock returns and find that context-aware networks outperform traditional methods, generating monthly seven-factor alphas of up to 281 basis points globally and 146 basis points in the U.S., with strong statistical significance. Importantly, these results are robust across portfolio weighting schemes, including equally weighted, value-weighted, and capped value-weighted, and persist after applying masking techniques to reduce look-ahead bias.

Second, we investigate the predictive power of our networks in identifying M&A targets. Building on Hoberg and Phillips (2010), who show that firms often acquire others with similar product offerings, we find that approximately half of U.S. acquisition targets fall within the top 100 most similar firms based on our network, comparable to the TNIC benchmark. Internationally, we observe similar patterns. A logistic regression confirms that textual similarity significantly increases the probability of an M&A event, even after controlling for observable firm characteristics. The predictive power decreases when using masked descriptions, indicating that some residual bias remains in unmasked networks, though this bias can be mitigated.

2.5 Aggregated Cluster Embeddings

In paper 4.5, I introduce Aggregated Cluster Embeddings (ACE), a novel approach that combines topic modeling with embedding models to semantically represent corporate risk disclosures. Using these representations, I show how to more accurately estimate factor betas such as market, size, value, profitability, and investment. The results demonstrate that a CatBoost model trained on ACE generates state-of-the-art beta estimations, particularly when combined with historical betas and other financial indicators. The method addresses the limitations of

traditional beta estimation approaches, which often assume time-invariant factor exposures and rely heavily on historical returns.

To construct ACE representations, I use a domain-specific embedding model trained on pre-2008 data to generate sentence embeddings from the risk disclosures in the Item 1A section of firms' 10-K reports. These sentences are then mapped to predefined risk clusters and their embeddings are pooled to form semantically rich, cluster-level vectors. These are concatenated to represent a firm's overall risk profile. Unlike previous studies that rely on word counts or topic frequencies, ACE captures qualitative differences in how firms describe risk. This is particularly important given that two firms can disclose the same risk category but in ways that imply opposite economic consequences.

The ACE framework has several advantages. First, it allows for interpretable beta estimation by tracing embeddings back to specific risk clusters, helping to mitigate the common "black box" problem associated with text-based machine learning models. Second, using a 2007 knowledge cutoff avoids look-ahead bias and enables truly out-of-sample beta forecasts beginning in 2008. Third, the approach is scalable and adaptable to international reports, earnings calls, and financial news, offering the potential for future applications beyond U.S. filings.

I illustrate the capabilities of ACE by using them as input to machine learning models that aim at estimating beta estimates. To estimate betas, I apply an expanding window approach, updating firm-level risk profiles annually while ensuring that all predictions are made on out-of-sample data. I use regularized and tree-based models (Ridge, Random Forest, and CatBoost) to account for the high dimensionality of the input vectors. Empirical results show that ACE-based models outperform standard rolling beta estimates and other benchmark models, including those using macroeconomic and technical indicators. For the market beta, estimation errors decline by around 20%. I also observe consistent improvements in size, value, profitability, and investment.

When combined with historical beta data and other financial features, ACE-based models achieve mean squared error (MSE) levels that rival or exceed the best-performing models in the literature. For example, while a baseline model following Drobetz et al. (2023) achieves an MSE of 0.1300, a combined model, incorporating ACE, technical indicators, and prior betas, achieves an MSE of 0.1244, a statistically significant improvement.

Beyond predictive performance, I assess which disclosed risks and linguistic features of the disclosures most influence beta estimates. Using explainable AI techniques, I find evidence suggesting that competition, supply chain disruptions, and regulatory uncertainty are the most

predictive risk factors, while intellectual property and goodwill impairment have weaker effects. Furthermore, I find that Disclosures framed in pessimistic language increase estimated market beta by over 10%, while those emphasizing long-term rather than short-term risks decrease beta by around 3.4%. Using terms like "major" or "likely" increases estimated beta values, suggesting that the content, tone, and modality of risk language affect investor perception. These effects are also consistent across other factors and are strongest in narrative sections where firms signal urgency or certainty.

3 Discussion

3.1 Academic Implications

This thesis makes several substantive contributions to the academic literature on asset pricing, investor behavior, and the integration of ML and NLP in financial economics.

First, it revisits the long-standing question of whether historical price data can predict future returns. By employing a random forest model trained on a comprehensive set of more than 180 technical indicators across 68 countries and several decades, this thesis provides new empirical evidence that cross-sectional return predictability persists globally. These indicators, which include trend-following rules, volatility oscillators, and volume-based signals, are widely used in practice yet remain largely underexplored in academic literature. Even though these indicators lack explicit theoretical foundations, the results demonstrate that they contain information that ML models can exploit to capture consistent patterns in the cross-section of returns. This challenges the notion that historical price patterns do not contain informational content and opens new avenues for theorizing the behavioral or structural mechanisms through which such predictability arises. Future research may build on this by examining whether the signals extracted by the ML model are consistent with behavioral biases, market microstructure frictions, or broader economic conditions.

Second, this thesis underscores the importance of target engineering in ML-based asset pricing. Rather than predicting continuous return values, which often result in a concentration on extreme observations, the model classifies whether a stock will outperform the median return in a given month. When training them for stock return prediction, this binary framing addresses the tendency of ML models to overweight illiquid or high-volatility stocks with extreme returns. By reducing sensitivity to outliers and allowing for a more robust learning process, this setup enhances the generalizability and stability of the predictions. Moreover, it lessens reliance on preprocessing techniques such as winsorization, which can inadvertently introduce researcher degrees of freedom and inflate performance through implicit overfitting.

Third, the thesis extends the literature on return predictability by demonstrating the utility of ML forecasts in portfolio construction, particularly within the classical mean-variance framework. Traditionally, expected returns are proxied using historical averages, a practice known to generate unstable portfolio weights due to estimation noise. This thesis introduces outperformance probabilities, derived from the classification model, as a bounded, forward-looking proxy for expected returns. When centered around zero, these probabilities fall within a compact interval [0.5,0.5], which reduces the influence of extreme values and yields more stable portfolio weights. Empirical evidence shows that portfolios constructed using these forecasts exhibit superior Sharpe ratios, reduced drawdowns, and improved diversification. This contribution addresses a gap in the academic literature by integrating ML-based signals into theoretically grounded optimization frameworks, bridging the divide between predictive modeling and classical portfolio theory.

Fourth, the thesis deepens our understanding of investor information processing by examining how market participants react to novel disclosures in corporate filings. Building on the literature that documents investor inattention to textual information (e.g., Cohen et al., 2020), this study extends the analysis to a global setting, demonstrating that inattention to novel textual content is a worldwide phenomenon, particularly pronounced in jurisdictions with stringent disclosure requirements. The newly introduced NID measure enables sentence-level identification of informational novelty by comparing current filings to past disclosures using contextual embeddings, allowing researchers to isolate new information in textual documents. This approach holds promise for broader applications, such as identifying overlooked content in analyst reports, filtering redundancy in financial news coverage, or calibrating attention-based investor models.

Fifth, the thesis contributes to the literature on the pricing of macroeconomic information by developing the ISI framework, which uses LLMs to quantify the impact of macroeconomic news on different industries. While previous studies often examine the aggregate effect of macroeconomic announcements, this approach introduces a scalable and interpretable method to assess differential exposure across sectors. This allows for more nuanced empirical tests of macro-news pricing and offers a new tool to explore how top-down shocks propagate through the cross-section of stocks. Moreover, the modular nature of the ISI framework suggests it could be extended to firm-level applications, where macroeconomic shocks are contextualized based on company-specific characteristics.

Sixth, this thesis advances our understanding of global inter-firm economic linkages by introducing a time-varying, global business network. While prior studies have developed time-varying business networks for U.S. firms (Hoberg and Phillips, 2010, 2016, 2025), these frameworks are limited in geographic scope and cannot capture global economic relationships. Leveraging AI-generated business descriptions from historical annual reports and computing the similarity of their contextualized embeddings, this thesis constructs the first comprehensive global business network spanning over 60,000 firms worldwide. This novel dataset enables the study of strategic interactions, competitive spillovers, and information diffusion at the international level. It provides a scalable foundation for future empirical research on cross-country economic dependencies, global asset pricing, trade linkages, and systemic risk propagation.

Finally, the thesis introduces ACE, a novel approach that integrates unsupervised topic modeling with embedding techniques to produce interpretable and domain-specific textual representations. Unlike standard sentence or document embeddings, which often lack interpretability, ACE maps semantic content into predefined, economically meaningful topics. Each dimension of the resulting vector corresponds to a distinct topic cluster, preserving interpretability while retaining the richness of high-dimensional embeddings. This balance is particularly valuable in asset pricing, where model explainability is critical for identifying the underlying drivers of risk and return. The thesis demonstrates the efficacy of ACE in estimating firm-specific factor betas and shows that variation in textual representations influences risk exposures in systematic ways. The ACE framework can be readily adapted to other research domains, such as ESG scoring, credit risk modeling, or corporate governance analysis.

3.2 Practical Implications

Beyond its academic contributions, this thesis provides a range of actionable insights for investors, corporate decision-makers, and policymakers. For investors, the demonstrated capability of ML models trained on technical indicators to predict relative stock performance offers valuable input for systematic portfolio construction. By focusing on relative returns rather than absolute levels, investors can avoid stocks with excessive volatility or illiquidity, leading to investment strategies that are more practical and implementable under real-world trading constraints. These findings are particularly relevant for institutional investors subject to liquidity regulations and retail investors operating under transaction cost limitations. Moreover, this thesis shows that integrating model-generated stock rankings into mean-variance optimization frameworks significantly enhances the out-of-sample performance of such portfolios in contrast

to backward-looking historical averages, especially during periods of heightened market volatility or structural shifts in the economic environment.

The insights on investor inattention to novel information in corporate disclosures also carry important practical implications. The NID approach introduced in this thesis enables investors to identify which parts of financial reports contain semantically new information by comparing textual embeddings across consecutive reports. This functionality supports more targeted reading and efficient information processing, helping investors focus on materially relevant content rather than repeatedly encountering redundant boilerplate language. Such filtering is especially valuable during earnings seasons or in environments where market participants must monitor a large number of disclosures within a short time frame. As a result, both institutional and retail investors can make more informed and timely decisions. At the same time, analysts and asset managers can allocate their attention more effectively to high-information-value firms or sectors.

Regarding macroeconomic information, the thesis demonstrates how industry impact scores derived from LLMs can assist investors in interpreting the implications of complex macroeconomic releases, such as those issued by central banks, multilateral institutions, or government agencies. These scores offer a systematic method for estimating macroeconomic information's directional and cross-sectional effects on different industries. They can be computed in near real-time using fine-tuned language models. The framework lends itself to automation and scalability, enabling integration into quantitative investment pipelines and macroeconomic monitoring dashboards. It could also be extended to analyze firm-specific news or peer disclosures. For instance, the same methodology could quantify the expected impact of regulatory changes, geopolitical events, or industry-level disruptions on individual firms, offering investors a more granular and timely basis for portfolio reallocation. Moreover, regulators could employ this tool to evaluate how timely macroeconomic signals are conveyed in official communications, improving transparency and market guidance.

The global business network developed in this thesis further supports practical decision-making by both investors and corporate managers. Investors may use the network to identify economically similar firms for peer comparison, performance benchmarking, or portfolio diversification, especially in international or emerging markets. Corporate decision-makers, in turn, can leverage the network to locate potential suppliers, strategic partners, or acquisition targets, particularly in regions where conventional business relationship data is limited or fragmented. The ability to distinguish competitors from supply chain partners using fine-tuned language

models increases the network's utility, as users can filter firms according to the specific type of economic link they are interested in exploring. This functionality can also aid multinational firms in increasing supply chain resilience, conducting scenario analysis, or identifying target firms for future M&A deals. Policymakers may also draw on this network to study the systemic implications of firm-level shocks and design resilience-enhancing policy interventions.

Finally, the ACE framework equips investors with a robust yet interpretable method for incorporating textual data into predictive models. Investors traditionally rely on structured, numerical data such as price-to-earnings ratios, balance sheet indicators, or technical signals to inform their investment decisions. ACE enables the integration of textual information, such as risk disclosures or management commentary, into machine learning pipelines while maintaining interpretability through topic-level representation. For instance, investors could use ACE to transform textual sections of annual reports into structured features, which can then be combined with technical and fundamental indicators to improve stock return predictions. Moreover, the topic-based decomposition makes it possible to trace which risk disclosures contribute most to the model's predictions, thereby offering valuable insights into which firm narratives the market tends to underreact to. This functionality may also be helpful for ESG-focused investors aiming to understand how risk communication aligns with a firm's sustainability or governance profile.

In summary, this thesis offers practical tools and frameworks that can be readily applied by financial market participants, enhancing decision-making processes, reducing informational inefficiencies, and supporting more adaptive responses to both firm-specific and macroeconomic developments.

3.3 Summary

This thesis leverages recent advances in ML and NLP to examine the efficiency with which investors process information in global capital markets. Across multiple empirical settings, the findings consistently reveal that investors fail to incorporate value-relevant signals embedded in both quantitative and qualitative data, resulting in persistent mispricings that challenge traditional asset pricing theories. By analyzing stock selection mechanisms, corporate disclosures, macroeconomic reports, and qualitative risk narratives, this research demonstrates the potential of AI-driven methods to enhance our understanding of international market efficiency.

The first paper introduces a machine learning–based stock ranking approach that emphasizes outperformance probabilities rather than raw return forecasts. This design mitigates common pitfalls in return prediction models, such as excessive focus on illiquid and volatile stocks. Empirical evidence shows that portfolios constructed using this probabilistic ranking framework generate substantial and statistically significant excess returns across regions, firm sizes, and market conditions. These returns are not fully explained by standard risk factors or by constraints to arbitrage. Furthermore, integrating outperformance probabilities into mean-variance optimization significantly reduces drawdowns and portfolio volatility, underscoring the robustness of such strategies, particularly in adverse market environments.

The second paper assesses whether investors promptly price newly disclosed information in international annual reports. A novel methodology is proposed to detect new content in disclosures by comparing sentence embeddings from current and prior reports using context-aware language models. This approach identifies semantically novel information that traditional text analysis methods often miss. In contrast to the predictions of the EMH, firms disclosing a higher volume of new information systematically underperform, while those with minimal novelty outperform. Long-short portfolios based on this information novelty yield highly significant alphas, indicating a systematic underreaction. Notably, the effect is strongest in jurisdictions with more stringent regulatory regimes, suggesting a greater proportion of value-relevant disclosures. These results highlight the limitations of investor attention and illustrate how advanced NLP techniques can uncover informational signals overlooked by conventional approaches.

The third paper investigates how investors process macroeconomic news, which often has heterogeneous effects across industries. We introduce a novel framework that leverages LLMs to generate ISI scores from macroeconomic reports issued by institutions such as the FED and the OECD. These scores quantify the expected directional impact of macroeconomic developments on specific industries. An investment strategy based on ISI scores, which takes long positions in industries expected to benefit and short positions in those likely to suffer, yields substantial and statistically significant abnormal returns. Specifically, the six-factor alpha reaches up to 93 basis points per month. This suggests a systematic underreaction to macroeconomic information by investors, even when the information is publicly available and economically meaningful. To rule out look-ahead bias, we conduct several diagnostic tests. We remove time-specific references from the reports before generating sentiment scores, apply out-of-sample prediction procedures, and analyze subdimension effects. The persistent profitability of the strategy, even after these precautions, indicates that the anomaly cannot be fully explained by look-ahead bias.

The fourth paper develops a time-varying global business network that captures firm-level relationships beyond standard industry classifications. While text-based networks have been studied in the context of the United States, applying similar methods to global markets is challenging due to variations in disclosure standards and reporting practices. This thesis addresses these limitations by constructing time-varying business descriptions and using context-aware embeddings to measure firm similarity. This approach uncovers operational and strategic connections often missed by conventional classification systems, thereby more accurately representing the growing interconnectedness of firms in the global economy. These findings demonstrate how AI and NLP can provide a more comprehensive understanding of corporate structures, offering valuable insights for asset pricing, mergers and acquisitions, and industrial organization.

The final paper presents ACE, a novel framework for integrating qualitative risk disclosures into asset pricing models. Conventional approaches to estimating factor betas typically rely on historical return correlations, which may overlook firm-specific risk exposures detailed in narrative reports. ACE addresses this limitation by applying topic modeling and sentence embeddings to transform textual risk disclosures from firms' annual reports into quantitative and interpretable representations. Empirical results demonstrate that a CatBoost model trained on ACE-based features significantly outperforms standard estimation techniques, particularly when combined with historical beta measures and traditional financial indicators. Among the identified risk categories, regulatory uncertainty, competitive pressure, and supply chain fragility emerge as especially influential in explaining variation in factor betas.

Overall, this thesis demonstrates that ML and NLP are powerful tools for identifying unpriced information across global capital markets. The empirical evidence challenges the notion of fully efficient markets by uncovering systematic underreactions to both qualitative and quantitative data. The results indicate that behavioral biases and structural limitations prevent investors from fully incorporating available information, leading to persistent mispricings that can be detected and exploited by machine learning–based approaches. While these advanced analytical methods reveal patterns often missed by traditional models, their widespread adoption may gradually diminish the anomalies they leverage, unless core frictions in information processing remain unresolved.

3.4 Limitations and Future Research

A key methodological challenge addressed in this thesis concerns the selection of suitable AI and LLMs. This choice often involves a trade-off between predictive accuracy and the risk of look-ahead bias. In general, more recent and better-trained models tend to perform more accurately. However, applying such models to historical data that predates their training period can inadvertently introduce forward-looking information, creating biased results. While this thesis employs several measures to reduce the impact of such bias, it often cannot be entirely ruled out. Future research should, therefore, prioritize the development of frameworks that balance predictive performance with temporal consistency and explicitly evaluate how model choice influences empirical findings.

Another critical issue for future research is replicability. Because Large Language Models (LLMs) are inherently non-deterministic, the same inputs may produce slightly different outputs across runs. This challenge is further complicated by the evolving nature of commercial models, which may be updated, replaced, or discontinued over time. These changes can limit the ability of researchers to replicate findings in the long term. To address this, future studies should examine how sensitive results are to variations in model choice, prompt design, and parameter settings. Clear documentation and extensive robustness checks will ensure that findings based on LLMs remain valid and credible.

Lastly, this thesis largely investigates qualitative and quantitative data separately. However, future research should aim to integrate both within a single analytical framework. The interaction between corporate narratives and firm-level financials may carry valuable information that is lost when the two are analyzed in isolation. Training AI models that integrate both qualitative and quantitative data, such as through the newly introduced ACE framework, holds the potential to provide a more comprehensive and accurate understanding of how markets process information.

Although ML and NLP tools offer powerful means of uncovering unpriced information, their broader impact on market efficiency is shaped by several factors. These include the speed of technological adoption, the regulatory environment, and the willingness of market participants to embrace data-driven analytical approaches. Future research should, therefore, examine the extent to which such models are already integrated into investment decision-making and explore the barriers that may prevent their widespread adoption. Understanding how investors engage

with these technologies will be essential for evaluating the long-term implications of this ongoing transformation in financial markets.

Bibliography

Andreou, P.C., Harris, T., Philip, D., 2020. Measuring firms' market orientation using textual analysis of 10-k filings. British Journal of Management 31, 872–895.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? the information content of internet stock message boards. The Journal of finance 59, 1259–1294.

Avramov, D., Cheng, S., Metzker, L., 2023. Machine learning vs. economic restrictions: Evidence from stock return predictability. Management Science 69, 2587–2619.

Boudoukh, J., Feldman, R., Kogan, S., Richardson, M., 2019. Information, trading, and volatility: Evidence from firm-specific news. The Review of Financial Studies 32, 992–1033.

Brown, S.V., Tucker, J.W., 2011. Large-sample evidence on firms' year-over-year md&a modifications. Journal of Accounting Research 49, 309–346.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Bybee, L., Kelly, B., Manela, A., Xiu, D., 2024. Business news and business cycles. The Journal of Finance 79, 3105–3147.

Campbell, J.L., Chen, H., Dhaliwal, D.S., Lu, H.m., Steele, L.B., 2014. The information content of mandatory risk factor disclosures in corporate filings. Review of Accounting Studies 19, 396–455.

Cao, S., Jiang, W., Yang, B., Zhang, A.L., 2023. How to talk when a machine is listening: Corporate disclosure in the age of ai. The Review of Financial Studies 36, 3603–3642.

Chen, Y., Kelly, B.T., Xiu, D., 2022. Expected returns and large language models. Available at SSRN 4416687 .

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2023. Palm: Scaling language modeling with

pathways. Journal of Machine Learning Research 24, 1–113.

Cochrane, J.H., 2011. Presidential address: Discount rates. Journal of Finance 66, 1047–1108.

Cohen, L., Malloy, C., Nguyen, Q., 2020. Lazy prices. The Journal of Finance 75, 1371–1415.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics. pp. 4171–4186.

Drobetz, W., Hollstein, F., Otto, T., Prokopczuk, M., 2023. Estimating stock market betas via machine learning. Journal of Financial and Quantitative Analysis , 1–56.

Fama, E.F., 1970. Efficient capital markets. Journal of finance 25, 383–417.

Fedyk, A., 2024. Front-page news: The effect of news positioning on financial markets. The Journal of Finance 79, 5–33.

Frankel, R., Jennings, J., Lee, J., 2022. Disclosure sentiment: Machine learning vs. dictionary methods. Management Science 68, 5514–5532.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Hoberg, G., Phillips, G., 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. The Review of Financial Studies 23, 3773–3811.

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of political economy 124, 1423–1465.

Hoberg, G., Phillips, G.M., 2025. Scope, scale, and concentration: The 21st-century firm. The Journal of Finance 80, 415–466.

Liu, Y., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 364.

Lopez-Lira, A., 2023. Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. Jacobs Levy Equity Management Center for Quantitative Financial Research Paper .

Lopez-Lira, A., Tang, Y., 2023. Can chatgpt forecast stock price movements? return predictability and large language models. arXiv preprint arXiv:2304.07619 .

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. The Journal of finance 66, 35–65.

Ludwig, J., Mullainathan, S., Rambachan, A., 2025. Large language models: An applied econometric framework. Technical Report. National Bureau of Economic Research.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 26.

OpenAI, 2023. GPT-4 technical report. arXiv:2303.08774.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Sarkar, S.K., Vafa, K., 2024. Lookahead bias in pretrained language models. Available at SSRN .

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of finance 62, 1139–1168.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: Quantifying language to measure firms' fundamentals. The journal of finance 63, 1437–1467.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems 30.

Appendix

4.1 Paper 1: Automated Stock Picking using Random Forests

Breitung, C. (2023)

Automated Stock Picking using Random Forests

Journal of Empirical Finance, Volume 72, Pages 532-556

https://doi.org/10.1016/j.jempfin.2023.05.001

VHB Journal Ranking 2025: B

Abstract

We derive a stock ranking by applying a technical features-based random forest model on an international dataset of liquid stocks. Rather than predicted return, our ranking is based on outperformance probability. By applying a decile split, we find that long–short portfolios achieve Sharpe ratios of up to 1.95 and a highly significant yearly six-factor alpha of up to 21.79%. Moreover, we show that outperformance probabilities serve as a superior measure of future returns in the context of portfolio optimization. Mean–variance portfolios using this measure are less volatile and more profitable than equally- or value-weighted portfolios. Our findings are robust to firm size, regional restrictions, and non-crisis periods and cannot be explained by limits to arbitrage.

Note:

The complete version of this published paper was provided in the examiner copies of this dissertation. To prevent issues related to plagiarism, duplicate publication, or copyright violation, it is not part of the publicly available version of the dissertation but can be accessed through the publisher's website.

4.2 Paper 2: New Information Detection using Language Models

Breitung, C. & Müller S. (2022)

New Information Detection using Language Models

Working Paper

Abstract

We introduce NID, a language model based approach to identify semantically new information within financial text. We showcase the performance of our measure by studying the *"lazy prices"* anomaly in an international setting using 500,000 annual reports from stocks across the globe. On the one hand, we find that NID identifies novel information with greater accuracy than traditional word-based techniques, generating a 89 bps monthly six-factor alpha. On the other hand, we obtain evidence that the anomaly primarily exists in markets with a strict regulatory oversight and predominantly affects smaller stocks. Notably, the results show no significant difference across low and high transaction costs, implying that limited attention, rather than limits to arbitrage, drives the observed anomaly.

Note:

The complete version of this working paper was provided in the examiner copies of this dissertation. To safeguard the integrity of the ongoing review process and to prevent plagiarism, duplicate publication, or copyright violations, the paper is not part of the publicly accessible version of this dissertation but is available upon request from the author.

4.3 Paper 3: Macroeconomic Inattention

Breitung, C., Kruthof G. & Müller S. (2025)

Macroeconomic Inattention

Working Paper

Abstract

We examine whether investors timely incorporate macroeconomic disclosures into asset prices. Leveraging LLMs, we develop a methodology to extract ISI scores from central bank and global institution reports. Using these scores, we construct long-short portfolios that generate highly significant six-factor alphas of up to 93 basis points per month, indicating a systematic underreaction to macroeconomic information. To ensure robustness, we conduct extensive tests for LLM-induced look-ahead bias and find no evidence in this regard. Additionally, we demonstrate how machine learning can enhance portfolio performance by optimizing the weighting of macroeconomic subdimensions. Our findings highlight the potential of LLM-driven text analysis in uncovering market inefficiencies and improving investment.

Note:

The complete version of this working paper was provided in the examiner copies of this dissertation. To safeguard the integrity of the ongoing review process and to prevent plagiarism, duplicate publication, or copyright violations, the paper is not part of the publicly accessible version of this dissertation but is available upon request from the author.

4.4 Paper 4: Global Business Networks

Breitung, C. & Müller S. (2025)

Global Business Networks

Journal of Financial Economics, April 2025, Article 104007

https://doi.org/10.1016/j.jfineco.2025.104007

VHB Journal Ranking 2025: A+

Abstract

We leverage the capabilities of GPT-3 to generate historical business descriptions for over 63,000 global firms. Utilizing these descriptions and advanced embedding models from OpenAI, we construct time-varying business networks that represent business links across the globe. We showcase the performance of these networks by studying the lead–lag effect for global stocks and predicting target firms in M&A deals. We demonstrate how masking firm-specific details can mitigate look-ahead bias concerns that may arise from the use of embedding models with a recent knowledge cutoff, and how to differentiate between competitor, supplier, and customer links by fine-tuning an open-source language model.

Note:

The complete version of this published paper was provided in the examiner copies of this dissertation. To prevent issues related to plagiarism, duplicate publication, or copyright violation, it is not part of the publicly available version of the dissertation but can be accessed through the publisher's website.

4.5 Paper 5: Aggregated Cluster Embeddings

Breitung, C. (2025)

Aggregated Cluster Embeddings

Working Paper

Abstract

This paper introduces ACE, a novel framework for semantically representing firms' disclosed risk factor information. By leveraging these representations, machine learning models trained with ACE achieve beta estimates with mean squared errors 20% lower than traditional OLS-based methods. Beyond improved accuracy, ACE provides qualitative insights into the risk factors and their presentation that may be used for factor beta estimation. When combined with historical betas and technical indicators, ACE-based models outperform state-of-the-art beta estimation methods while maintaining interpretability. Notably, these results are free from look-ahead bias, as ACE employs a finance-specific embedding model with a 2007 knowledge cutoff.

Note:

The complete version of this working paper was provided in the examiner copies of this dissertation. To safeguard the integrity of the ongoing review process and to prevent plagiarism, duplicate publication, or copyright violations, the paper is not part of the publicly accessible version of this dissertation but is available upon request from the author.