# Classifying Hedge Fund Strategies with Large Language Models: Systematic vs. Discretionary Performance

Hui-Ching Chuang, Chung-Ming Kuan

*Department of Statistics, National Taipei University*

*Department of Finance, National Taiwan University*

**Abstract**

This paper fine-tuned the FinBERT, a large language model (LLM) tailored for the financial domain, to classify hedge funds into Systematic and Discretionary categories. By leveraging LLM techniques, our approach mitigates the subjective judgment traditionally involved in categorizing investment strategies. We find that on average, funds classified as Systematic yield higher factor-adjusted returns than their Discretionary counterparts. Moreover, after implementing test with a false discovery adjustment, we observe that between 10% to 20% of funds exhibit statistically significant positive alphas in models combining of observable and unobservable factors.

*Keywords:* Transfer learning; Textual analysis; False discovery rate; Fund performance
*JEL:* C63; G11; G14; G23

*We would like to express our appreciation for the constructive comments and valuable insights provided by anonymous referees, Jules Van Binsbergen, Dacheng Xiu, Po-Hsuan Hsu, O-Chia Chuang, and the participants of the Quantitative Finance Workshop 3: Asset Pricing and Risk Management, IMS, Singapore; and the 26th Conference on the Theories and Practices of Securities and Financial Markets.

**Corresponding author: Hui-Ching Chuang. Address: Department of Statistics, National Taipei University. No. 151, University Rd., Sanxia Dist., New Taipei City, Taiwan 237. Phone: 886-9727-35021. We gratefully acknowledge the financial support from the National Science and Technology Council, Taiwan. Version Date: 2025/04/20.

*Email addresses:* hcchuang@gm.ntpu.edu.tw (Hui-Ching Chuang), ckuan@ntu.edu.tw (Chung-Ming Kuan)

## 1. Introduction

Investment strategies are complex decision processes involving quantitative and qualitative market information assessments. Such strategies play a crucial role in hedge fund performance. A fund's investment strategy is usually disclosed in its private placement memorandum or fund prospectus. Fund managers tend to avoid specific descriptions of their strategies to permit investment flexibility. As such, it is not straightforward to categorize hedge funds based on their disclosed statements. On the other hand, the Hedge Fund Research (HFR) database offers its fund classification systems; for example, HFR strategy classification in 2017 includes five major categories: Equity Hedge, Event-Driven, Macro, Relative Value and Fund of Funds, each with several sub-strategy groups. While the HFR classification provides valuable information about the fund characteristics, much effort is still needed if one would like to categorize hedge funds under different criteria.

With the advancement of analytical tools and computational technology, more fund managers now rely on models, algorithms, and various learning methods to make investment decisions. Thus, it would be interesting to classify hedge funds into "systematic" and "discretionary" funds and study how these two groups of funds perform in practice. By systematic funds, we refer to the funds with strategies depending mainly on quantitative models without human intervention by discretionary funds we refer to those require primarily managers' professional skills and experience. Such classification is also in line with the HFR categorization for the sub-categories of Macro funds: Systematic Diversified funds and Discretionary Thematic funds.[1] Similarly, hedge/mutual funds are classified as "man" and "machine" in Harvey et al. (2017) and "quantitative" and "discretionary" ("non-quantitative") in Abis (2022) (Beggs and Hill-Kleespie, 2025), or "quantitative" and "fundamental" in Evans, Rohleder, Tentesch and Wilkens (2023).

In this paper, we introduce an approach for building classifiers that bifurcate hedge funds into systematic and discretionary categories, leveraging the large language model (LLM). Specifically, we extract features from Systematic Diversified and Discretionary Thematic hedge funds to construct a training sample, and fine-tune FinBERT, a BERT-based model tailored for the financial

[1] HFR defines Systematic Diversified funds as funds with "investment processes that typically are functions of mathematical, algorithmic and technical models, with little or no influence from individuals over the portfolio positioning," and Discretionary Thematic funds are those "primarily reliant on the evaluation of market data, relationships, and influences, as interpreted by an individual or group of individuals who make decisions on portfolio positions."

domain, to accommodate the content of fund strategy descriptions. BERT (Bidirectional Encoder Representations from Transformers), proposed by Devlin et al. (2019), learns contextual information from both the left and right sides of a word, and is pre-trained on a large general corpus including Wikipedia and BookCorpus. FinBERT, developed by Yang et al. (2020) and Huang et al. (2023), further train the model on SEC corporate filings (10-K and 10-Q), financial analyst reports from Thomson Investext, and earnings call transcripts from SeekingAlpha. As a result, FinBERT captures contextual nuances in financial texts than the original BERT model. Our classifier architecture is based on FinBERT, with a classification head added to distinguish hedge fund styles. The resulting model effectively leverages financial language understanding to differentiate between systematic and discretionary strategies. Our classification task is similar to that applied by Abis (2022), Beggs and Hill-Kleespie (2025), and Harvey et al. (2017), differing mainly in that we rely on guidance from HFR's internal expert classification as training labels and do not require selecting keywords to label the funds.

We evaluate fund performance using the FDR-based test proposed by Giglio et al. (2021) to test multiple alphas in the linear asset pricing models. We then compare whether positive alpha funds (we use outperforming funds for positive alpha funds interchangeably) belong to the systematic funds (or discretionary funds) and the magnitude of their performance difference. Examining the positiveness of thousands of individual funds' alpha is a multiple-testing question. Multiple testing is easy to suffer from the data-snooping bias, i.e., we are likely to identify the outperforming funds purely due to chance. The false discovery fallacy is critical when searching for positive alpha funds. Benjamini and Hochberg (1995) (BH) are pioneers proposing the test to examine the multiple hypotheses while controlling the false discovery rate (FDR), which is defined as the expected value of the false rejected number to the rejected number of the hypotheses. FDR and related multiple testing approaches have attracted more and more finance researchers recently, see, e.g., Harvey et al. (2020), Chordia et al. (2020), Hsu et al. (2024) and others.[2]

Under the conventional Fama-MacBeth two-pass regression framework, Giglio et al. (2021) propose a rigorous multiple-test framework that accommodates missing data and omitted risk

[2]Barras et al. (2010), Cuthbertson et al. (2012), Bajgrowicz and Scaillet (2012), Bajgrowicz et al. (2016) use the Bayesian FDR control test. Another strand of literature focuses on controlling the family-wise error rate (FWER), which is the probability of committing more than one false discovery, see White (2000), Hansen (2005), Romano and Wolf (2005), and Hsu and Kuan (2005). The applications of FWER control include the profitability of trading strategies: Kuang et al. (2014), Goyal and Wahal (2015). Harvey et al. (2016) and Chordia et al. (2020) also promote the FDP, false discovery proportion introduced in Romano and Wolf (2007), and Romano et al. (2008).

factors. Hedge fund data is known for its short life span (unbalanced panel return structure, missing values), herding trade (cross-sectional dependence), and highly nonlinear payoff structures (possibly the existence of latent risk factors). These characteristics and the generated variable bias from the two-pass procedure threaten the underlying independence assumptions of Benjamini and Hochberg (1995) test. Giglio et al. (2021) propose the adjustment to the conventional two-pass methods and FDR test, which mitigate those issues' threat to the validity of BH test and further improve the power of the test while maintaining the FDR control.

Our classification task includes the two sub-strategies of the Macro fund (training sample) and four sub-strategies of the Equity Hedge funds (testing sample) in the HFR database. We fine-tuning the FinBERT model to 85% of the training and validation sample and compare the prediction results on the 15% hold out samples with respect to various bag-of-words based machine learning approaches. Results show that the FinBERT model has the highest hold-out sample prediction ability, yields as high as 93% , 96%, 92%, and 95% in terms of accuracy, area under the ROC curve, precision, and F1 scores.

We study the performance of those classified funds surviving at least 36 months from 1994 to 2015. We find that systematic funds have higher Sharpe ratios and factor-adjusted alphas than those classified as discretionary. These results hold for all four sub-strategies of the Equity Hedge funds and are robust to one, three, five, and seven risk factor models. The FDR-based multiple alphas test shows that there are 10% to 20% statistically significant positive alpha funds in both catogories.

Our research makes several contributions to the literature. First, we propose a novel approach to classify the style of hedge fund investing strategies. The quantitative (non-quantitative) investment style of funds draws the researcher's attention to its impact on the market liquidity, tail risk, the economy of scale, and others (Abis, 2022; Evans et al., 2023). Our fine-tuned FinBERT approach helps extract the textual information from the funds with well-defined classification styles and predict the less clearly defined styles of funds. It reduces researcher-dependent judgment effort while keeping the classification consistent with well-defined styles.

Second, we identify the proportion of the authentic outperforming funds in systematic and discretionary funds using the test without data-snooping bias. Our results add to the research of Giglio et al. (2021) on the performance of Hedge fund's style investment and also give rigorous

statistical evidence on the profitability of the systematic and discretionary funds in addition to Chincarini (2014), and Harvey et al. (2017).

Third, we show that fine-tuning a content-specific LLM using fund prospectuses achieves high classification performance. Domain-specific LLMs have demonstrated strong performance in areas such as science and biomedicine (Beltagy et al., 2019; Lee et al., 2020), legal studies (Chalkidis et al., 2020), ESG research (Huang et al., 2023; Webersinke et al., 2021), and innovation studies (Lee and Hsiang, 2020; Chuang et al., 2023). Our analysis extends this promising line of work by showing that LLMs can automatically assign hedge fund styles, helping to decode the potentially complex strategies employed across the fund industry.

This paper proceeds as follows. Section 2 discusses the methods for extracting features from fund strategy descriptions and for building fund classifiers from these features. In Section 3, we evaluate the performance of the classified systematic and discretionary funds and compare their overall performance via FDR tests. The last section concludes the paper.

## 2. Classification of Hedge Funds

In this section, we discuss our approach to fine-tuning FinBERT to classify systematic and discretionary funds, based on the documents of fund investment strategies. Following Harvey et al. (2017), we consider the two largest groups in the HFR classification system: Macro funds and Equity Hedge funds, where the former includes two sub-strategy groups (Systematic Diversified funds and Discretionary Thematic funds), and the latter contains four sub-strategy groups (Equity Market Neutral funds, Fundamental Growth funds, Fundamental Value funds, and Quantitative Directional funds).[3] Given that Macro funds have already been classified into systematic and discretionary funds, it is quite natural to use the information of Macro funds to train classifiers. All strategy descriptions are sourced from the HFR database; we include the graveyard database to mitigate survivorship bias. In total, we collect 2,242 Macro-fund strategy descriptions: 1,479 classified as systematic and 763 as discretionary.

Fine-tuning the FinBERT model of Yang et al. (2020); Huang et al. (2023) proceeds in three stages: tokenization and embedding, transformer encoding and pooling, and the final classification step. Tokenization converts raw text into a sequence of token IDs based on FinVocab, inserts

[3] As Harvey et al. (2017), we ignore sector-specific funds and those with "multistrategy".

special tokens (`[CLS]` at the start and `[SEP]` to separate sentences), and pads or truncates to a fixed length of 512 tokens.[4] Each token ID is then mapped to a $1 \times 768$ token embeddings vector (a vector of real numbers to be estimated), capturing both semantics and position information in the sequence.

The second stage, transformer encoding and pooling, refines the initial token embeddings by incorporating contextual information from the entire input sequence and outputs a representative vector. Specifically, the $512 \times 768$ token embedding matrix is fed into a 12-layer, bidirectional $\text{BERT}_{\text{BASE}}$ Transformer encoder (encoder comprises weight matrices to be estimated). At each layer, the self-attention mechanism acts as an adaptive weighting scheme, updating each token's (latent) factor loadings based on its association with all other tokens in both its left and right context (see Vaswani et al. (2017)). After twelve iterations, the encoder yields 768-dimensional loadings, or termed transformer embeddings, one per token (including the special `[CLS]` token). Finally, a pooling layer applies a linear transformation and a Tanh activation function to the `[CLS]` vector, yielding a single 768-dimensional summary representation of the raw text, which serves as the input to the final classification task.

The final classification stage applies a linear transformation to the summary vector, yielding two-dimension logits. (i.e., by multiplying the summary vector by a $2 \times 768$ weight matrix plus 2-vector constant). A softmax function converts these logits into the predicted probabilities for two fund styles.

In our training process, we withheld 15% of the Macro strategy descriptions (337) as a hold-out test set. From the remaining 85%, we reserved 15% (286) for validation to monitor model performance during training, leaving 1,619 descriptions for training. To maintain the original 65.97% systematic–discretionary ratio, we employed stratified sampling. The pre-trained FinBERT model is loaded from the Hugging Face platform,[5] We then fine-tuned all model parameters using our strategy descriptions of the training sample. Detailed hyper-parameter settings in the training

[4]Tokens may represent words, subwords, or mixture with punctuation. For example, the sentence *"Fine-tuning the FinBERT model rocks."* is split into WordPiece tokens `['fine', '-', 'tuni', '##ng', 'the', 'fin', '##bert', 'model', 'rock', '##s', '.']` with corresponding token IDs `[3, 4882, 30861, 16256, 1071, 6, 3388, 16909, 674, 5102, 63]`. FinVocab is the list of financial-related tokens comprising 30,873 case-insensitive entries, each occurring at least 8,500 times in a corpus of: (i) 2.5 billion tokens from Russell 3000 firms' Form 10-K/10-Q filings (business descriptions, risk factors, MD&A, 1994–2019); (ii) 1.1 billion tokens from S&P 500 analyst reports (2003–2012); and (iii) 1.3 billion tokens from earnings-call transcripts of 7,740 public firms (2004–2019). Two special tokens, `[CLS]`, denoting the beginning and `[SEP]` to separate the sentence pair are inserted into the token list.

[5]The model is listed on `https://huggingface.co/yiyanghkust/finbert-pretrain`

architecture, and epoch (training rounds) performance are provided in Appendix A.

Our classification performance is listed in Table 1. We show classification performance metrics for train, validation, and test samples using the fine-tuned FinBERT model. Results show that our fund-strategy-fine-tuned FinBERT model has high hold-out sample prediction ability, yields as high as 93%, 96%, 92%, and 95%, 97% in terms of accuracy, AUC, precision, F1, and recall scores. These scores are generally higher than those obtained from standard machine learning methods, with improvements ranging from 1 to 10 percentage points. To provide additional context, Appendix A.II outlines the methodological details and performance results of several bag-of-words-based classifiers.[6]

Figure 1 illustrates the distribution and decile breakdown of predicted probabilities for funds classified under the Macro strategy. The upper panel overlaid a histogram of predicted probabilities with a kernel density estimate. The bimodal shape indicates strong model confidence, with most funds assigned probabilities close to either 0 or 1. The lower panel groups Macro funds into ten deciles based on their predicted probabilities. Within each decile, we compute the proportion of funds labeled as Systematic Diversified or Discretionary Thematic, following the visualization style of Brachtendorf et al. (2023). The diverging bar plot reveals clear alignment between predicted probabilities and true labels: Discretionary funds dominate the lower deciles, while Systematic funds concentrate on the upper ones. Notably, only a limited number of funds fall within the intermediate range (roughly between the 0.47 and 0.93 decile groups), where model predictions are less decisive. Overall, Figure 1 demonstrates that the predicted probabilities effectively bifurcate Macro funds along their actual strategy types.

We then extend the fine-tuned FinBERT model to classify Equity Hedge funds. To interpret classification outcomes, we construct ranked bigram tables (two consecutive words) ranked table for both Macro and Equity Hedge funds, grouped by predicted style. Table 2 shows the most frequent bigrams associated with each predicted style within the two strategy categories. To enhance the informativeness, we remove globally common phrases (top 5%) and apply a standard set of text preprocessing procedures: lowercasing, stopword and punctuation removal, Porter stemming, and part-of-speech filtering to retain only nouns and proper nouns. Figure 2 visualizes these results as

[6]Due to slight differences in sample construction between the bag-of-words ML classifiers and the FinBERT fine-tuning setup (align the bi-grams used in both main strategies, remove punctuations, etc.,), the results presented in the Appendix are not intended to be directly comparable.

word clouds.

Across both Macro and Equity Hedge funds, predicted Systematic strategies exhibit recurring technical terms such as algorithm, model, and comput, reflecting their rules-based approach. In contrast, predicted Discretionary funds are characterized by terms related as market, report, and research. These lexical distinctions affirm that our model captures meaningful semantic differences between systematic and discretionary fund strategies.

## 3. Comparison of Fund Performance

We assess the performance of these funds by comparing their excess returns, Sharpe ratios, and alphas derived from various factor models and across the different categorized groups. In line with prior studies on hedge funds, our analysis is limited to funds that report monthly returns and adopt the "Net of All Fees" reporting style, and delete the first 12 return observation to avoid the back-filled biasas, seen in Cao et al. (2013). Furthermore, we narrowed our dataset to include only those funds with at least 36 consecutive monthly returns to ensure robust regression outcomes. The final dataset for our performance evaluation consists of 3,905 Equity Hedge funds and 1,129 Macro funds from January 1994 to November 2015.

### *3.1. Performance Based on Excess Returns and Alphas*

We consider the following factor model:

$$E(r_i) = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{\lambda}, \quad i = 1, \cdots, N, \tag{1}$$

where $r_i$ is the excess return of fund $i$ (excess of the risk-free rate), $\alpha_i$ is the pricing error (alpha) of fund $i$, $\boldsymbol{\beta}_i$ is the $S \times 1$ vector of risk exposures to $S$ risk factors, and $\boldsymbol{\lambda}$ is the $S \times 1$ vector factor risk premia (reward for risk exposure), and $N$ is the number of funds; see, e.g., Cochrane (2009). A fund is considered superior if its alpha is greater than zero, suggesting positive abnormal return.

In this study, we opted for models with 1, 3, 5, 7, and 11 factors, denoted as F1,F3,F5, F7, and F11, respectively. For the 1-factor model, the only risk factor considered is the market factor (MKT), calculated as the value-weighted return of all CRSP firms in excess of the risk-free rate. For the 3-factor model, the risk factors include MKT, SMB (small minus big), and HML (high minus low). SMB and HML represent size and book-to-market equity mimicking portfolios in stock

returns, as defined by Fama and French. For the 5-factor model, according to Fung and Hsieh (2001, 2004)Fung and Hsieh (2001, 2004), the risk factors are PTFSBD, PTFSFX, PTFSCOM, PTFSIR, and PTFSSTK. These factors represent the returns from the long position of the lookback straddle of bonds, currencies, commodities, short-term interest rates, and stocks. Finally, for the 7-factor model, we consider MKT, SMB, CS (credit spread), $\Delta$10Y, PTFSBD, PTFSFX, and PTFSCOM as risk factors. CS is the monthly change in the difference between a BAA bond yield and a 10-year constant maturity Treasury yield (GS10). $\Delta$10Y represents the long-term interest rate, specifically the monthly change of GS10. The 11-factor model adds additional four factors to the 7-factor model: HML, MOM, PTFSIR, and PTFSSTK, where MOM is the momentum factor of Carhart (1997).[7] Simple time series regression for each funds are performed to obtained the intercept as fund alpha.

Table 3 summaries excess-return moments, Sharpe ratios, skewness, and kurtosis for the full fund sample, the two main strategy: Equity Hedge and Macro; and the four Equity Hedge sub-strategies, all separated by the Discretionary and Systematic styles predicted by our fine-tuned FinBERT classifier. The sample tilts toward Discretionary funds overall (3,066 versus 1,968 Systematic); the same pattern holds within Equity Hedge (2,728 versus 1,177) but reverses in Macro (338 versus 791). Although mean excess returns are nearly identical across styles—45.6% for Discretionary and 45.0% for Systematic—their distributions differ markedly: Discretionary funds display wider cross-sectional dispersion and noticeably fatter tails, as reflected in higher standard deviations, more negative skewness, and greater kurtosis. These contrasts are strongest in the Quantitative Directional segment, where FinBERT classifies 89 funds as Discretionary and 171 as Systematic, highlighting the model's ability to separate judgement-driven from algorithmic trading.[8] Figure 3 depicts the excess-return and Sharpe-ratio distributions for Macro and Equity Hedge funds. In concert with Table 3, the figure demonstrates that, while average performance is broadly comparable across styles, Discretionary and Systematic funds differ markedly in volatility, skewness, and tail behavior.

[7]Risk-free rate and factors MKT, SMB, and HML are sourced from Kenneth R. French's website. The five hedge fund factors can be found on Professor David A. Hsieh's website: http://faculty.fuqua.duke.edu/~dah7. Data for BAA and GS10 are available through the Federal Reserve Economic Data of the Federal Reserve Bank of St. Louis. Table A.2 in the appendix summarizes the risk factors' means, medians, standard deviations, minimums, and maximums and risk factors' correlation matrix.

[8]Quantitative Directional strategies, by design, exploit statistical and factor models to extract predictive patterns from historical prices, an approach consistent with the stronger systematic representation.

To control for market-wide and style-specific risk factors, Table 4 next reports the distribution of alpha estimates for Systematic versus Discretionary funds—by main and sub-strategy—using the 1-, 3-, 5-, 7-, and 11-factor models, together with the associated mean differences. After adjusting for factor exposures, the performance gap between systematic and discretionary styles generally widens. Equity Hedge systematic strategies deliver alphas that exceed their discretionary peers by approximately 5.4 pp under the market model (F1), 6.3 pp under the Fama–French three-factor model (F3), 7.3 pp under the seven-factor model (F7), and 5.0 pp under the eleven-factor model (F11); the sole exception is the Fung–Hsieh five-factor model (F5), where systematic funds underperform by about 0.9 pp. Macro systematic funds demonstrate even larger alpha advantages outperforming discretionary by roughly 24.1 pp in F1, 21.2 pp in F3, 31.5 pp in F5, 27.0 pp in F7, and 25.9 pp in F11, underscoring the robust excess returns captured by systematic approaches once factor risks are accounted for.

Table 5 shows that Systematic funds outperform Discretionary peers in three of the four Equity Hedge sub-strategies, Equity Value, Fundamental Growth and Fundamental Value, by roughly 4 to 8 pp of alpha across factor models (peaking under the eleven-factor specification). The biggest gap is in Quantitative Directional, where Systematic strategies outperform Discretionary by over 10–19 pp, underscoring the alpha-generating power of statistical directional trading once common risks are removed.

### *3.2. Significant Performance under False Discovery Rate Control*

In addition to comparing the factor returns, we also delve into the statistical significance of each fund's performance. To achieve this, we formulate the following multiple hypotheses:

$$H_{0,i} : \alpha_i \leq 0, \quad i = 1, \cdots, N. \tag{2}$$

Refuting the null hypothesis $H_{0,i}$ implies that the superior performance (positive alpha) of fund $i$ is statistically significant and cannot merely be attributed to chance. Instead, it may indicate the fund manager's genuine investment acumen.

We used the test from Giglio et al. (2021) to identify funds with positive alpha in each category. The test by Giglio et al. (2021) includes several steps. First, they use observable risk factors to calculate risk exposures and residuals for each fund through time-series regression. Second, they

employ matrix completion on the unbalanced residual matrix, Hastie et al. (2015), and use PCA to identify latent risk factors and exposures. Then, they perform a cross-sectional regression of the mean excess return on the concatenated observed and unobserved exposures to estimate risk premiums and fund alphas. To account for potential estimation errors in alpha, the alpha estimates are debiased before applying the alpha-screening B-H test, a power-enhanced version of the original B-H test that accounts for inequality in hypotheses. Detailed information on the alpha estimation algorithm and alpha-screening B-H test can be found in Append A.III.[9]

Table 6 displays the ratio of rejected hypotheses (i.e., funds with positive alpha) to the total number of hedge funds, categorized accordingly. The FDR is maintained below the 5% level. Columns (3) to (6) represent results from observable 3-, 5-, 7-, and 11-factor models, referred to as Fs models where $s$ equals 3, 5, and 7. Columns (7) and (8) combine a 7-factor model with 4-, 2-unobservable factors to 3 and 5 observables, respectively, labeled as F3+U4 and F5+U2. The final column consider pure unobservable factor models with 7 factors, denoted as U7. Panel A outlines the proportion of positive alpha funds within discretionary or systematic funds across all funds, and Panel B lists results on Macro and Equity Hedge, while Panel C focuses on the four sub-strategies of Equity Hedge funds. The proportion of positive alpha is defined as the significant number within each classified style as main or sub-strategy considered. For example, considering the F3 model in column (3) in Panel A, 21.33% refers to there are 21.33% of all 3,066 classified Discretionary funds' F3 alpha is positive significantly, and 20.12% of all 1,177 classified Systematic funds' F3 alpha is positive significantly.

Results in Table 6 highlight distinct style effects at the strategy level. In Panel A, the overall share of significant alphas is slightly lower or comparable for Systematic funds (20.12% in F3, 14.43% in F5, and 18.70% in F7) versus Discretionary funds (21.33% in F3, 14.12% in F5, and 22.02% in F7). When broken down by main strategy in Panel B, a more nuanced pattern emerges: within Equity Hedge, Systematic funds lead (24.55% vs. 22.84% in F7), whereas in Macro, Discretionary funds prevail (15.38% vs. 9.99% in F7). These results also holds for F11 and mixture of F3+U4, F5+U2, and U7 factor models. These results indicate that the Systematic style excels in Equity Hedge, whereas the Discretionary style yields more significant positive alphas in Macro.

[9]The alpha-screening FDR test was conducted using a program developed by Giglio et al. (2021), available at https://dachxiu.chicagobooth.edu/.

## 4. Conclusions

This paper studies hedge fund classification and performance and contributes to the hedge fund literature. First, we introduce a large language model approach, fine-tuning FinBERT, to classifying hedge funds into systematic and discretionary funds that differ from existing methods. Second, we use the false discovery control test to examine whether factor-adjusted returns (alphas) of the classified Systematic and Discretionary funds' performance. Our empirical results show that Systematic funds are preferred to Discretionary funds across all categories of funds we considered in terms of factor adjusted returns (alphas). We identify 10% to 20% of the authentic positive alpha funds while controlling for the multiple test bias.

## References


Abis, S., 2022. Man vs. machine: Quantitative and discretionary equity management. Working Paper .

Bajgrowicz, P., Scaillet, O., 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. Journal of Financial Economics 106, 473–491.

Bajgrowicz, P., Scaillet, O., Treccani, A., 2016. Jumps in high-frequency data: Spurious detections, dynamics, and news. Management Science 62, 2198–2217.

Barras, L., Scaillet, O., Wermers, R., 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. The Journal of Finance 65, 179–216.

Beggs, W., Hill-Kleespie, A., 2025. Quantitative investing and market instability: Evidence from mutual fund fire sales. Available at SSRN 3281447 .

Beltagy, I., Lo, K., Cohan, A., 2019. Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 .

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

Brachtendorf, L., Gaessler, F., Harhoff, D., 2023. Truly standard-essential patents? A semantics-based analysis. Journal of Economics & Management Strategy 32, 132–157.

Cao, C., Chen, Y., Liang, B., Lo, A.W., 2013. Can hedge funds time market liquidity? Journal of Financial Economics 109, 493–516.

Carhart, M.M., 1997. On persistence in mutual fund performance. The Journal of Finance 52, 57–82.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I., 2020. LEGAL-BERT: The muppets straight out of law school. arXiv preprint arXiv:2010.02559 .

Chincarini, L., 2014. The impact of quantitative methods on hedge fund performance. European Financial Management 20, 857–890.

Chordia, T., Goyal, A., Saretto, A., 2020. Anomalies and false rejections. The Review of Financial Studies 33, 2134–2179.

Chuang, H.C., Hsu, P.H., Lee, Y.N., Walsh, J.P., 2023. What share of patents is commercialized? Technical Report. Working Paper. Manuscript.

Cochrane, J.H., 2009. Asset Pricing: Revised edition. Princeton University Press.

Cuthbertson, K., Nitzsche, D., O'Sullivan, N., 2012. False discoveries in UK mutual fund performance. European Financial Management 18, 444–463.

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), pp. 4171–4186.

Evans, R.B., Rohleder, M., Tentesch, H., Wilkens, M., 2023. Diseconomies of scale in quantitative and fundamental investment styles. Journal of Financial and Quantitative Analysis 58, 2417–2445.

Fung, W., Hsieh, D.A., 2001. The risk in hedge fund strategies: Theory and evidence from trend followers. The Review of Financial Studies 14, 313–341.

Fung, W., Hsieh, D.A., 2004. Hedge fund benchmarks: A risk-based approach. Financial Analysts Journal 60, 65–80.

Giglio, S., Liao, Y., Xiu, D., 2021. Thousands of alpha tests. The Review of Financial Studies 34, 3456–3496.

Goyal, A., Wahal, S., 2015. Is momentum an echo? Journal of Financial and Quantitative Analysis 50, 1237–1267.

Hansen, P.R., 2005. A test for superior predictive ability. Journal of Business & Economic Statistics 23, 365–380.

Harvey, C.R., Liu, Y., Saretto, A., 2020. An evaluation of alternative multiple testing methods for finance applications. The Review of Asset Pricing Studies 10, 199–248.

Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. The Review of Financial Studies 29, 5–68.

Harvey, C.R., Rattray, S., Sinclair, A., Van Hemert, O., 2017. Man vs. machine: Comparing discretionary and systematic hedge fund performance. The Journal of Portfolio Management 43, 55–69.

Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity: the Lasso and Generalizations. CRC press.

Hsu, P.H., Kuan, C.M., 2005. Reexamining the profitability of technical analysis with data snooping checks. Journal of Financial Econometrics 3, 606–628.

Hsu, P.H., Kyriakou, I., Ma, T., Sermpinis, G., 2024. Mutual funds' conditional performance free of data snooping bias. Journal of Financial and Quantitative Analysis , 1–28.

Huang, A.H., Wang, H., Yang, Y., 2023. Finbert: A large language model for extracting information from financial text. Contemporary Accounting Research 40, 806–841.

Kuang, P., Schröder, M., Wang, Q., 2014. Illusory profitability of technical analysis in emerging foreign exchange markets. International Journal of Forecasting 30, 192–205.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36, 1234–1240.

Lee, J.S., Hsiang, J., 2020. Patent classification by fine-tuning bert language model. World Patent Information 61, 101965.

Romano, J.P., Shaikh, A.M., Wolf, M., 2008. Formalized data snooping based on generalized error rates. Econometric Theory 24, 404–447.

Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73, 1237–1282.

Romano, J.P., Wolf, M., 2007. Control of generalized error rates in multiple testing. The Annals of Statistics 35, 1378–1408.

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.

Webersinke, N., Kraus, M., Bingler, J.A., Leippold, M., 2021. Climatebert: A pretrained language model for climate-related text. arXiv preprint arXiv:2110.12010 .

White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.

Yang, Y., Uy, M.C.S., Huang, A., 2020. Finbert: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097 .

Figure 1: Distribution and decile breakdown of predicted probability within Macro strategy. The upper panel presents the kernel density estimate of the predicted probability of Systematic, and the lower panel shows the relative proportions of systematic and discretionary funds across each predicted decile group.

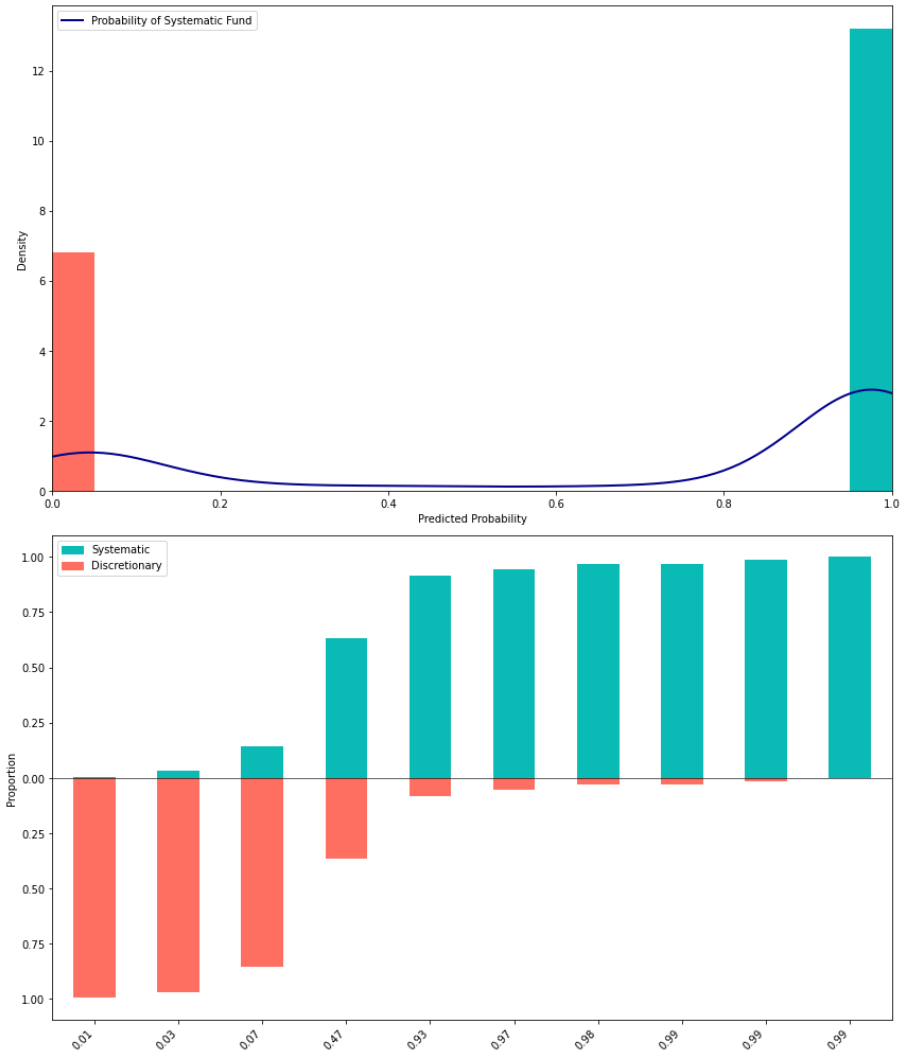
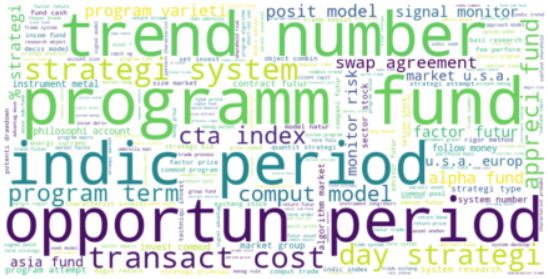


Figure 2: Predicted style and bigram-cloud. This plot presents the top-ranked bigrams (two consecutive words) extracted from strategy descriptions for each predicted style within the Equity Hedge and Macro categories. Bigrams are selected based on frequency after excluding globally common phrases (top 5%), and applying text preprocessing steps including lowercasing, stopword and punctuation removal, Porter stemming, and part-of-speech filtering to retain only nouns and proper nouns.


(a) Equity Hedge: Discretionary


(b) Equity Hedge: Systematic


(c) Macro: Discretionary




(d) Macro: Systematic

Figure 3: Style, excess returns, and Sharpe ratios: 1994-2015.

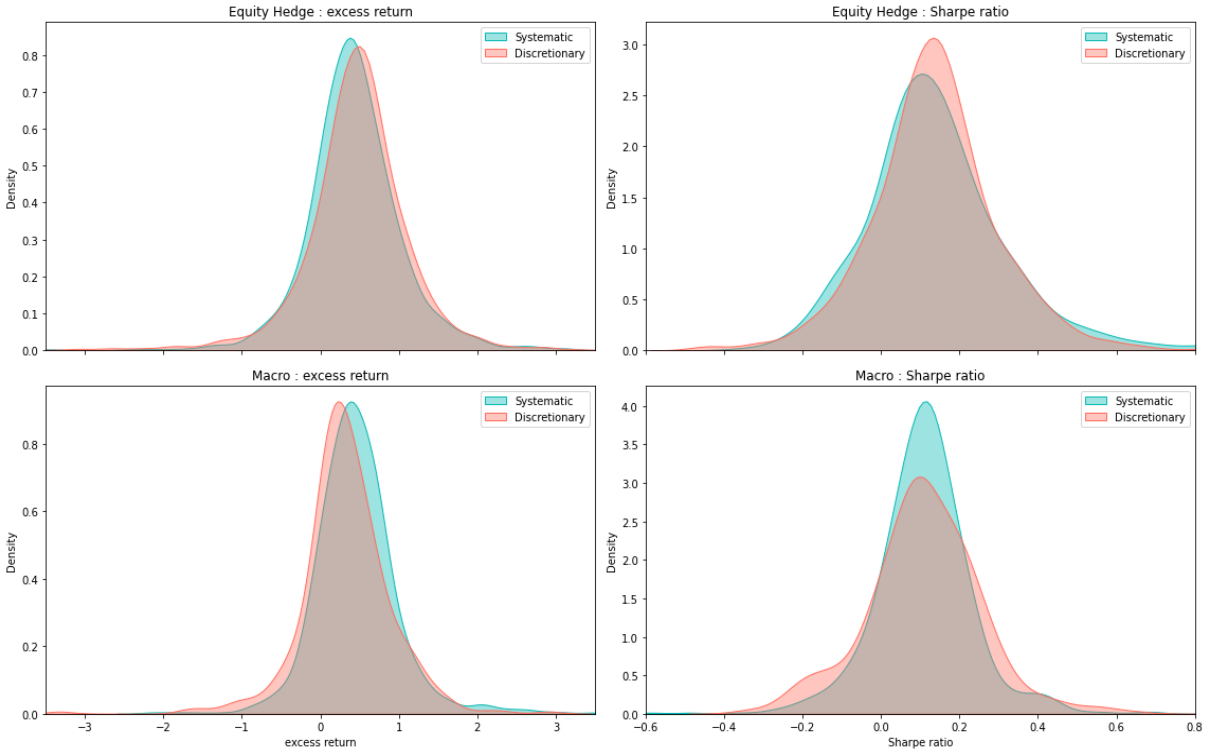


Table 1: Classification performance measures

| | Accuracy | AUC | Precision | F1 | Recall |
|---|---|---|---|---|---|
| Training | 93.14% | 97.55% | 92.38% | 94.95% | 97.66% |
| Validation | 89.16% | 93.47% | 89.90% | 91.99% | 94.18% |
| Test | 92.58% | 96.16% | 91.91% | 94.53% | 97.30% |

This table reports the classification performance of the fine-tuned FinBERT model. The model is based on the pretrained model by Huang et al. (2023) and fine-tuned on a stratified Macro fund sample comprising 1,619 training observations, 286 validation observations, and 337 hold-out test observations.

Table 2: Predicted style and top-ranked bigrams by Equity Hedge and Macro strategy

| Equity Hedge | | Macro | |
|---|---|---|---|
| Discretionary | Systematic | Discretionary | Systematic |
| zar fund | liquid screen | flow analysi | programm fund |
| burden market | algorithm model | invest target | opportun period |
| structur return | europ eastern | restrict respect | trend number |
| broker research | europ wace | latin america | indic period |
| cap bia | model strategi | basi sub-fund | strategi system |
| economi compani | comput algorithm | exampl posit | transact cost |
| issuer japan | sector neutral | secur repres | day strategi |
| cycl industri | combin index | issu invest | program term |
| bond govern | select factor | futur basi | cta index |
| market russia | equiti model | strategi master | appreci fund |
| relationship compani | return number | analysi currenc | comput model |
| invest and/or | methodolog fund | invest idea | posit model |
| sub-fund medium | appreci tokyo | bond incom | swap agreement |
| state ‘ci | section topix | asia pacif | program varieti |

This table presents the top-ranked bigrams (two consecutive words) extracted from strategy descriptions for each predicted style within the Equity Hedge and Macro categories. Bigrams are selected based on frequency after excluding globally common phrases (top 5%), and applying text preprocessing steps including lowercasing, stopword and punctuation removal, Porter stemming, and part-of-speech filtering to retain only nouns and proper nouns.

Table 3: Summary statistics of excess return, Sharpe ratio, skewness, and kurtosis

| Panel A: Excess Returns and Sharpe Ratios | | | Excess return | | | Sharpe ratio | | |
|---|---|---|---|---|---|---|---|---|
| | | Count | Mean | STD | 50% | Mean | STD | 50% |
| Panel A.1: All Strategy | | | | | | | | |
| | Discretionary | 3,066 | 45.59% | 68.16% | 46.38% | 13.63% | 16.99% | 13.09% |
| | Systematic | 1,968 | 44.95% | 59.71% | 40.47% | 13.17% | 17.75% | 11.85% |
| Panel A.2: Main Strategy | | | | | | | | |
| Equity Hedge | Discretionary | 2,728 | 47.23% | 69.17% | 48.51% | 13.95% | 17.16% | 13.42% |
| | Systematic | 1,177 | 44.20% | 62.60% | 40.39% | 14.64% | 20.02% | 12.49% |
| Macro | Discretionary | 338 | 32.33% | 57.78% | 28.11% | 11.06% | 15.36% | 10.35% |
| | Systematic | 791 | 46.06% | 55.14% | 41.55% | 10.98% | 13.39% | 10.90% |
| Panel A.3: Sub-strategies of Equity Hedge | | | | | | | | |
| Equity Market Neutral | Discretionary | 202 | 25.27% | 43.15% | 20.04% | 14.15% | 21.59% | 10.81% |
| | Systematic | 385 | 29.90% | 42.02% | 27.84% | 17.65% | 26.55% | 13.37% |
| Fundamental Growth | Discretionary | 1,091 | 46.88% | 77.47% | 49.70% | 12.43% | 16.76% | 12.03% |
| | Systematic | 203 | 56.06% | 77.17% | 49.41% | 11.99% | 14.18% | 12.13% |
| Fundamental Value | Discretionary | 1,346 | 50.21% | 64.99% | 50.98% | 15.29% | 16.91% | 14.68% |
| | Systematic | 418 | 49.84% | 64.70% | 49.55% | 13.81% | 15.95% | 13.18% |
| Quantitative Directional | Discretionary | 89 | 56.38% | 61.56% | 52.33% | 11.99% | 12.25% | 12.20% |
| | Systematic | 171 | 48.56% | 71.07% | 40.47% | 13.03% | 16.90% | 10.99% |
| Panel B: Skewness and Kurtosis | | | Skewness | | | Kurtosis | | |
| | | Count | Mean | STD | 50% | Mean | STD | 50% |
| Panel B.1: All Strategy | | | | | | | | |
| | Discretionary | 3,066 | -14.61% | 92.24% | -16.41% | 266.98% | 449.86% | 142.27% |
| | Systematic | 1,968 | -7.08% | 94.84% | -2.68% | 246.32% | 529.33% | 117.03% |
| Panel B.2: Main Strategy | | | | | | | | |
| Equity Hedge | Discretionary | 2,728 | -16.49% | 88.23% | -17.81% | 255.46% | 392.66% | 143.32% |
| | Systematic | 1,177 | -22.31% | 103.48% | -21.57% | 306.63% | 634.17% | 155.28% |
| Macro | Discretionary | 338 | 0.61% | 118.86% | 0.69% | 359.89% | 763.72% | 136.61% |
| | Systematic | 791 | 15.58% | 74.79% | 15.25% | 156.59% | 292.37% | 69.92% |
| Panel B.3: Sub-strategies of Equity Hedge | | | | | | | | |
| Equity Market Neutral | Discretionary | 202 | -17.83% | 90.80% | -5.03% | 277.88% | 371.06% | 144.67% |
| | Systematic | 385 | -27.16% | 94.20% | -19.95% | 273.92% | 449.23% | 117.07% |
| Fundamental Growth | Discretionary | 1,091 | -18.60% | 87.28% | -18.89% | 246.83% | 398.46% | 129.74% |
| | Systematic | 203 | -20.35% | 76.32% | -28.55% | 234.88% | 281.54% | 155.28% |
| Fundamental Value | Discretionary | 1,346 | -14.34% | 89.10% | -17.59% | 261.77% | 396.09% | 153.40% |
| | Systematic | 418 | -21.29% | 121.75% | -16.54% | 388.53% | 920.29% | 190.27% |
| Quantitative Directional | Discretionary | 89 | -20.11% | 80.94% | -18.53% | 215.11% | 307.44% | 111.07% |
| | Systematic | 171 | -16.17% | 103.24% | -31.62% | 265.21% | 357.50% | 140.26% |

Table 4: Factor alphas by predicted style within Equity Hedge and Macro strategy

| Main Strategy | Style | Count | Mean | STD | 25% | 50% | 75% | Mean Diff. |
|---|---|---|---|---|---|---|---|---|
| Panel A: Market model (F1) | | | | | | | | |
| Equity Hedge | Discretionary | 2,728 | 13.79% | 73.70% | -12.11% | 19.24% | 48.08% | 5.39% |
| | Systematic | 1,177 | 19.18% | 62.96% | -11.12% | 19.16% | 47.53% | |
| Macro | Discretionary | 338 | 18.35% | 62.83% | -6.40% | 16.87% | 50.06% | 24.14% |
| | Systematic | 791 | 42.49% | 59.93% | 13.21% | 41.50% | 71.67% | |
| Panel B: Fama-French three-factor model (F3) | | | | | | | | |
| Equity Hedge | Discretionary | 2,728 | 9.62% | 73.64% | -15.40% | 15.56% | 43.28% | 6.31% |
| | Systematic | 1,177 | 15.93% | 64.10% | -15.95% | 15.92% | 44.66% | |
| Macro | Discretionary | 338 | 15.77% | 62.61% | -9.40% | 15.05% | 48.00% | 21.18% |
| | Systematic | 791 | 36.95% | 63.08% | 5.22% | 34.44% | 65.91% | |
| Panel C: Fung-Hsieh Five-factor model (F5) | | | | | | | | |
| Equity Hedge | Discretionary | 2,728 | 34.17% | 77.43% | 2.09% | 35.63% | 68.77% | -0.86% |
| | Systematic | 1,177 | 33.31% | 72.07% | 0.18% | 32.07% | 65.95% | |
| Macro | Discretionary | 338 | 30.63% | 66.58% | 3.61% | 24.08% | 61.29% | 31.48% |
| | Systematic | 791 | 62.11% | 72.80% | 24.95% | 58.34% | 94.65% | |
| Panel D: Seven-factor model (F7) | | | | | | | | |
| Equity Hedge | Discretionary | 2,728 | 8.24% | 77.80% | -20.41% | 15.29% | 45.52% | 7.28% |
| | Systematic | 1,177 | 15.52% | 68.54% | -15.40% | 15.82% | 46.52% | |
| Macro | Discretionary | 338 | 10.20% | 67.83% | -13.13% | 7.52% | 42.87% | 26.99% |
| | Systematic | 791 | 37.19% | 69.46% | 1.25% | 33.57% | 70.01% | |
| Panel E: Eleven-factor model (F11) | | | | | | | | |
| Equity Hedge | Discretionary | 2,728 | 11.35% | 84.83% | -20.50% | 17.10% | 51.09% | 5.02% |
| | Systematic | 1,177 | 16.37% | 76.27% | -17.15% | 17.40% | 51.25% | |
| Macro | Discretionary | 338 | 13.86% | 77.85% | -18.95% | 15.59% | 49.17% | 25.92% |
| | Systematic | 791 | 39.79% | 80.08% | -6.39% | 39.39% | 79.12% | |

This table reports factor alphas across predicted styles (Discretionary vs. Systematic) within the Equity Hedge and Macro strategies, estimated under five factor models. "'Mean Diff" refers to the mean difference between Discretionary and Systematic styles within each Main-strategy.

Table 5: Factor alphas by predicted style within four sub-strategies of Equity Hedge

| Sub Strategy | Style | Count | Mean | STD | 25% | 50% | 75% | Mean Diff |
|---|---|---|---|---|---|---|---|---|
| Panel A: Market model (F1) | | | | | | | | |
| Equity Market Neutral | Discretionary | 202 | 20.40% | 49.64% | -2.65% | 14.61% | 42.23% | 2.85% |
| | Systematic | 385 | 23.25% | 40.86% | 0.17% | 22.05% | 46.84% | |
| Fundamental Growth | Discretionary | 1,091 | 1.99% | 86.18% | -25.40% | 12.42% | 44.40% | 8.54% |
| | Systematic | 203 | 10.53% | 79.30% | -28.68% | 12.00% | 46.69% | |
| Fundamental Value | Discretionary | 1,346 | 22.50% | 65.46% | -2.65% | 25.54% | 51.72% | -3.58% |
| | Systematic | 418 | 18.92% | 64.91% | -14.20% | 19.08% | 49.08% | |
| Quantitative Directional | Discretionary | 89 | 11.70% | 47.01% | -18.29% | 8.30% | 34.36% | 9.20% |
| | Systematic | 171 | 20.90% | 75.61% | -11.68% | 18.17% | 46.13% | |
| Panel B: Fama-French three-factor model (F3) | | | | | | | | |
| Equity Market Neutral | Discretionary | 202 | 18.97% | 48.31% | -4.11% | 12.13% | 38.72% | 2.47% |
| | Systematic | 385 | 21.44% | 41.26% | -4.07% | 20.87% | 45.27% | |
| Fundamental Growth | Discretionary | 1,091 | -3.13% | 86.33% | -35.54% | 9.73% | 39.16% | 9.08% |
| | Systematic | 203 | 5.95% | 80.33% | -31.78% | 9.93% | 43.42% | |
| Fundamental Value | Discretionary | 1,346 | 18.65% | 65.06% | -6.24% | 21.11% | 46.78% | -4.16% |
| | Systematic | 418 | 14.49% | 66.94% | -21.15% | 13.18% | 41.03% | |
| Quantitative Directional | Discretionary | 89 | 7.94% | 47.23% | -20.52% | 7.02% | 29.43% | 10.91% |
| | Systematic | 171 | 18.86% | 75.62% | -17.57% | 14.92% | 48.90% | |
| Panel C: Fung-Hsieh Five-factor model (F5) | | | | | | | | |
| Equity Market Neutral | Discretionary | 202 | 26.98% | 46.50% | 1.16% | 20.73% | 48.02% | 2.94% |
| | Systematic | 385 | 29.92% | 42.68% | 4.81% | 24.51% | 53.53% | |
| Fundamental Growth | Discretionary | 1,091 | 30.10% | 87.11% | -2.61% | 35.40% | 71.10% | 1.67% |
| | Systematic | 203 | 31.77% | 81.15% | -8.47% | 34.92% | 80.32% | |
| Fundamental Value | Discretionary | 1,346 | 38.50% | 72.17% | 8.04% | 37.93% | 70.08% | -5.61% |
| | Systematic | 418 | 32.89% | 75.47% | -4.81% | 36.08% | 66.73% | |
| Quantitative Directional | Discretionary | 89 | 34.89% | 81.45% | -9.79% | 29.96% | 61.24% | 8.92% |
| | Systematic | 171 | 43.81% | 99.39% | -0.50% | 37.97% | 82.77% | |
| Panel D: Seven-factor model (F7) | | | | | | | | |
| Equity Market Neutral | Discretionary | 202 | 18.68% | 49.60% | -3.18% | 17.06% | 39.00% | 3.17% |
| | Systematic | 385 | 21.85% | 43.13% | -6.32% | 21.43% | 47.59% | |
| Fundamental Growth | Discretionary | 1,091 | -5.17% | 90.58% | -38.56% | 7.67% | 41.94% | 5.25% |
| | Systematic | 203 | 0.08% | 83.07% | -32.38% | 1.45% | 39.00% | |
| Fundamental Value | Discretionary | 1,346 | 17.87% | 68.56% | -11.54% | 20.68% | 49.06% | -3.50% |
| | Systematic | 418 | 14.37% | 70.46% | -18.24% | 14.49% | 45.86% | |
| Quantitative Directional | Discretionary | 89 | 3.26% | 67.39% | -22.25% | 4.40% | 47.56% | 19.12% |
| | Systematic | 171 | 22.38% | 86.57% | -9.76% | 23.54% | 48.82% | |
| Panel E: Eleven-factor model (F11) | | | | | | | | |
| Equity Market Neutral | Discretionary | 202 | 19.57% | 54.78% | -3.55% | 17.18% | 44.67% | 2.54% |
| | Systematic | 385 | 22.10% | 45.68% | -3.37% | 21.18% | 47.74% | |
| Fundamental Growth | Discretionary | 1,091 | -2.59% | 99.00% | -42.30% | 7.17% | 47.33% | 1.20% |
| | Systematic | 203 | -1.39% | 98.99% | -39.15% | -2.32% | 42.16% | |
| Fundamental Value | Discretionary | 1,346 | 21.67% | 74.61% | -9.16% | 24.27% | 53.63% | -6.26% |
| | Systematic | 418 | 15.41% | 73.15% | -20.63% | 14.33% | 53.32% | |
| Quantitative Directional | Discretionary | 89 | 7.65% | 73.73% | -27.05% | 5.29% | 47.46% | 19.24% |
| | Systematic | 171 | 26.89% | 100.97% | -11.86% | 26.34% | 61.45% | |

This table reports factor alphas across predicted styles (Discretionary vs. Systematic) within four sub-strategies of Equity Hedge, estimated under five different factor models. "Mean Diff" refers to the mean difference between Discretionary and Systematic styles within each sub-strategy.

Table 6: Proportion of significant alphas by style across strategies and sub-strategies

| | Style | F3 | F5 | F7 | F11 | F3+U4 | F5+U2 | U7 |
|---|---|---|---|---|---|---|---|---|
| Panel A: All strategies | | | | | | | | |
| | Discretionary | 21.33% | 14.12% | 22.02% | 19.57% | 22.64% | 23.39% | 22.11% |
| | Systematic | 20.12% | 14.43% | 18.70% | 17.28% | 21.39% | 20.93% | 19.21% |
| Panel B: Main strategies | | | | | | | | |
| Equity Hedge | Discretionary | 22.10% | 14.44% | 22.84% | 20.16% | 23.17% | 24.05% | 22.69% |
| | Systematic | 25.40% | 17.59% | 24.55% | 22.94% | 24.38% | 25.66% | 23.53% |
| Macro | Discretionary | 15.09% | 11.54% | 15.38% | 14.79% | 18.34% | 18.05% | 17.46% |
| | Systematic | 12.26% | 9.73% | 9.99% | 8.85% | 16.94% | 13.91% | 12.77% |
| Panel C: Sub-strategies of Equity Hedge | | | | | | | | |
| Equity Market Neutral | Discretionary | 23.76% | 19.80% | 26.73% | 21.78% | 23.27% | 23.76% | 26.73% |
| | Systematic | 28.83% | 24.42% | 29.87% | 28.83% | 27.53% | 30.65% | 28.05% |
| Fundamental Growth | Discretionary | 16.87% | 11.64% | 18.52% | 15.67% | 19.25% | 21.17% | 19.43% |
| | Systematic | 16.75% | 7.39% | 13.79% | 12.81% | 15.76% | 24.14% | 17.73% |
| Fundamental Value | Discretionary | 27.04% | 16.57% | 26.45% | 24.22% | 26.89% | 26.82% | 25.26% |
| | Systematic | 27.75% | 17.46% | 26.32% | 24.64% | 27.27% | 23.92% | 24.88% |
| Quantitative Directional | Discretionary | 7.87% | 4.49% | 12.36% | 10.11% | 14.61% | 17.98% | 14.61% |
| | Systematic | 22.22% | 14.62% | 21.05% | 17.54% | 20.47% | 20.47% | 16.96% |

This table reports the proportion of significant positive alphas by predicted style across different strategy levels. Panel A presents results for all strategies, Panel B reports by main strategy (Equity Hedge and Macro), and Panel C focuses on sub-strategies within Equity Hedge. Significant positive alphas are identified using the false discovery rate (FDR) procedure of Giglio et al. (2021), controlling the FDR at the 5% level. Fs denotes an observable s-factor model, where s = 3, 5, 7, or 11. Uk denotes an unobserved k-factor model. We also include hybrid models that combine observable and unobservable components: F3+U4 and F5+U2, as well as a fully latent 7-factor model, U7.

## Appendix

### A.I. FinBERT and fine-tuning procedures

Figure A.1 illustrates the architecture of the pre-trained FinBERT model, including the classification head for two fund styles. The associated size for the learning parameters are also displayed in the figure. To fine-tuning the model, We first randomly sample 20 learning rates from a logarithmic range between 1e-6 and 5e-5 to determine the optimal setting. We then train the model using 10 training epochs, a batch size of 8, and a weight decay of 0.01. The grid search and fine-tuning procedures closely follow the implementation guidelines provided by the FinBERT[1] and Hugging Face Transformers documentation.[2] Codes are available upon request.

Figure A.2 visualizes the training process and performance evaluation. The upper panel displays training and evaluation loss over epochs, while the lower panel plots validation metrics, including accuracy, AUC, precision, and F1 score, across training epochs.

### A.II. Machine leaning text classification comparison

We first follow standard practice in textual analysis to process the text of strategy descriptions. We exclude digital numbers, punctuation, symbols, and the stop-words (e.g., is, at, and, the) in all documents that are of little value for classification. The remaining words are then lemmatized, i.e., different forms of a word is converted to one single word, from which documents are tokenized based on "bigrams" (two consecutive words). To ensure a bigram in Macro funds (the training sample) is also relevant in Equity Hedge funds; we set the ratio of the percentage of Equity Hedge funds with a particular bigram to the percentage of Macro funds with the same bigram to be greater than or equal to 0.2.

We then construct the feature matrix of a given fund category as follows. For the token $j$ in the document $i$, its "term frequency" (tf) is:

$$\text{tf}_{ij} = \frac{\text{Number of times that token } j \text{ appear in the document } i}{\text{Total number of all tokens in the document } i}, \quad i = 1, \ldots, N, j = 1, \ldots, M,$$

and every tf is weighted by the inverse-document frequency (idf):

$$\text{idf}_j = \log \frac{\text{Total number of documents}}{\text{Number of the documents that contain token } j}, \quad j = 1, \ldots, M,$$

where $N$ is the number of funds in a category (Macro funds or Equity Hedge funds), and $M$ is the number of tokens. Note that the larger the idf, the less frequently the token $j$ is observed in these documents; such token is considered more informative for classification and hence receives more weight. The feature matrix is an $N \times M$ matrix with the $(i, j)$-th element:

$$f_{ij} = \text{tf}_{ij} \cdot \text{idf}_j, \quad i = 1, \ldots, N, \ j = 1, \ldots, M.$$

In our study, the feature matrix of Macro funds is $2,222 \times 3,494$, and that of Equity Hedge funds is $7,158 \times 3,494$.

We define the binary target variable as taking the value 1 if it is a Systematic Diversified fund and 0 if it is a Discretionary Thematic fund and the training sample are the feature matrix of Macro funds. The following statistical learning methods are employed: Linear regression, logistic regression, linear discriminant analysis (LDA), $k$-nearest neighbor (KNN), support vector machine (SVN) with the Gaussian kernel, classification tree, bagging, gradient boosting, as well as random forests. Our training approach utilizes text mining and statistical learning and

[1] https://github.com/yya518/FinBERT/blob/master/finetune.ipynb
[2] https://huggingface.co/docs/transformers/en/hpo_train

hence avoids subjectivity to a large extent. Moreover, the use of bigrams for tokenization also alleviates the problem of misinterpreting single words.

We select the trained classifier with the best classification performance. To this end, we consider four performance measures: Accuracy, area under the receiver operating characteristic curve (AUC), precision, and F1 score. We evaluate the performance of classifiers using the nested 10-fold cross-validation. This cross-validation involves two layers: the inner 10-fold cross-validation determines the best hyper-parameters for each learning method, and the outer 10-fold cross-validation evaluates the classification ability of different classifiers with the best hyper-parameters. In our study, we split Macro funds into two sub-samples, one with 85% (1,888) funds and the other with 15% (334) funds, by the stratified sampling on the strata of the Systematic Diversified dummy. The nested 10-fold cross-validation is then applied to the sub-sample of 85% Macro funds to search for the best classifier; the remaining 15% of Macro funds is reserved for out-of-sample evaluation of the best classifier.

Table A.1 contains four panels, where each panel summarizes the performance results of all learning methods under a particular measure. We report the summary statistics (median, mean, maximum, minimum, and standard deviation) based on the outer 10-fold samples. It can be seen that random forest dominates other classifiers for all measures in terms of these statistics, with gradient boosting as the second-best classifier. On the other hand, linear regression, logistic regression, and LDA perform pretty poorly. For example, the mean of accuracy is 0.86 for random forest and 0.84 for gradient boosting. On the other hand, linear regression, logistic regression, and LDA have respective means of 0.70, 0.70, and 0.68. Applying the selected random forest with the best hyper-parameters to the 15% validation sample, the resulting out-of-sample accuracy, AUC, precision, and F1 score are, respectively, 0.89, 0.87, 0.90, and 0.92, which are all greater than the corresponding medians and means in Table A.1.

## A.III. False-discovery adjusted procedures

This appendix follows the Algorithm 6 and 7 in Giglio et al. (2021).

### *A.III.1. Estimate alpha under the unbalanced panel and observable and unobservable mixture factor models*

Assume the general factor model with $S$ observable factors and $K$ unobservable factors:

$$E(r_i) = \alpha_i + \boldsymbol{\beta}'_{i,o}\boldsymbol{\lambda}_o + \boldsymbol{\beta}'_{i,u}\boldsymbol{\lambda}_u, \quad i = 1, \cdots, N,$$

where $\boldsymbol{\beta}_{i,o}$ and $\boldsymbol{\beta}_{i,u}$ are $S \times 1$ and $K \times 1$ risk exposure to the observable and unobservable risk factors. $\boldsymbol{\lambda}_o$ and $\boldsymbol{\lambda}_u$ are the risk premium of the asset for bearing observable and unobservable risks respectively. Assume that the excess return of fund $i$ at time $t$ is $r_{i,t}$, $i = 1, \cdots, N$; and $t \in \mathcal{T}_i$, which is the time indices set which of fund $i$ has excess return. $\mathcal{N}_t$ is the fund's indices set which includes the existing funds at time $t$.

Step 1. Time series regression. For each fund, estimate the time-series regression of excess return on the observable risk factors with the same range to obtain the observable risk exposure $\hat{\boldsymbol{\beta}}_{i,o}$ and residual $e_{i,t}$ for $t \in \mathcal{T}_i$. Let $E_{N\times T}$ be the residual matrix (with missing values).

Step 2. Matrix completion of the residual matrix. Suppose $E = M + U$, where $M$ is a $N \times T$ low rank matrix, and $U$ is the noise. Let $\Omega$ indicate the existing status of the matrix $E$, i.e., $\omega_{i,t} = 1$ if $e_{i,t}$ is observed, and 0 if missing. The projection matrix, $P_\Omega(E)$ imputes zeros on the missing entries of matrix $E$ as

$$\left[P_\Omega(E)\right]_{i,t} = \begin{cases} e_{i,t}, & \text{if } \omega_{i,t} = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{C.1}$$

We want to find a low-rank matrix, $M$, such that minimizes the

$$\min_{M \in R^{N \times T}} \left\| (E - M) \circ \Omega \right\|_F^2 + c\|M\|_*,$$

where $\circ$ is the element-wise product of matrices; $c$ is the tuning parameter; $\|M\|_F$ is the Frobenius norm. $\|M\|_F^2 := \sqrt{\sum_i \sum_t |m_{i,t}|^2}$; and $\|M\|_*$ is the nuclear norm. $\|M\|_* := \sum_{j=1}^{\min\{N,T\}} \sigma_j(M)$, where $\sigma_1(M) \geq \sigma_2(M) \geq \cdots$ are the ordered singular values of $M$. The iterative approach to obtain estimates of $M$, $\hat{M}$, see Hastie et al. (2015) and Giglio et al. (2021).

Step 3. Unobservable factors and exposure estimate. Apply singular value decomposition on the matrix $\hat{M}$, and define the unobservable $K \times 1$ factors and their exposures as:

$$\hat{f}_{u,t} = \left( \sum_{i \in \mathcal{N}_t} u_i u_i' \right)^{-1} \sum_{i \in \mathcal{N}_t} u_i e_{i,t},\ t = 1, \cdots, T,$$

$$\hat{\boldsymbol{\beta}}_{u,i} = \left( \sum_{t \in \mathcal{T}_i} \hat{f}_{u,t} \hat{f}_{u,t}' \right)^{-1} \sum_{t \in \mathcal{T}_i} \hat{f}_{u,t} e_{i,t},\ i = 1, \cdots, N,$$

where $u_1, \cdots, u_K$ is the top $K$ left singular-vector of $\hat{M}$. Define all risk exposures as $\hat{\boldsymbol{\beta}} := (\hat{\boldsymbol{\beta}}_o, \hat{\boldsymbol{\beta}}_u)$ and all observable and unobservable risk factor as $\hat{f}_t := \left( f_{o,t} - \bar{f}_o, \hat{f}_{u,t}' \right)'$, where $f_{o,t}$ is the observable $S \times 1$ risk factors for $t = 1, \cdots, T$, and $\bar{f}_o = \frac{1}{T} \sum_{t=1}^{T} f_{o,t}$.

Step 4. Estimate risk premium. Run a cross-section regression of $\bar{r}_i$ on $\hat{\boldsymbol{\beta}}$ to obtain the slope $\hat{\boldsymbol{\lambda}}$ as the risk premium.

Step 5. De-biased alpha estimates.

$$\hat{\alpha}_i = \bar{r}_i - \hat{\boldsymbol{\beta}}_i' \hat{\boldsymbol{\lambda}} + \hat{A}_i.$$

where $\hat{A}_i$ is the (de-)biased term for the unbalanced data, see Giglio et al. (2021).

Step 6. Construct the t-statistics and its $p$-values. The $t$-statistics is the standard asymptotic normal for one-side test.

$$t_i = \frac{\hat{\alpha}_i}{se(\hat{\alpha}_i)}, \quad p_i = 1 - \Phi(t_i), \quad i = 1, \cdots, N,$$

$\Phi(\cdot)$ is the standard normal CDF, and $se(\hat{\alpha}_i) = \frac{1}{|\mathcal{T}_i|} \sqrt{\sum_{t \in \mathcal{T}_i} \hat{\epsilon}_{i,t}^2 \left( 1 - \hat{f}_t' \hat{\Sigma}_f^{-1} \hat{\boldsymbol{\lambda}} \right)^2}$, where $\hat{\epsilon}_{i,t} = r_{i,t} - \bar{r}_i - \hat{\boldsymbol{\beta}}_i' \hat{f}_t$, $\hat{\Sigma}_f = \frac{1}{T} \sum_{t=1}^{T} \hat{f}_t \hat{f}_t'$.

*A.III.2. Alpha-screening Benjamini and Hochberg (1995) FDR control*

It is well known that simultaneously testing multiple hypotheses is easy to suffer from the false discovery problem. Suppose $t_i$ is a test statistics to examine $H_{0,i}$ in equation (**??**). A null hypothesis is rejected when $t_i > c_i$ for a threshold $c_i$. Let $\mathcal{H}_0$ is the set of indices of the true null hypotheses, $\mathcal{R}$ is the set of indices of the rejected hypotheses, and $\mathcal{F}$ is the indices of false rejected hypotheses, i.e.,

$$\mathcal{R} = \bigcup_{1 \leq i \leq N} \{ i : t_i > c_i \};$$
$$\mathcal{F} = \bigcup_{1 \leq i \leq N} \{ i : t_i > c_i \text{ and } \alpha_i \leq 0 \}.$$

The false discovery proportion is defined as the number of falsely rejected hypotheses to the total number of rejections. As the number of false rejection is unobservable, the false discovery rate (FDR) is then defined as the expectation of false discovery proportion, i.e.,

$$FDR := E\left(\frac{|\mathcal{F}|}{|\mathcal{R}|}\right).$$

where $|A|$ denotes the number of elements in the set $A$. If the number of rejections is zero, then FDR is defined as zero. Benjamini and Hochberg (1995) proposed the following procedures to control the FDR under $q$ level. Let

$$p_{(1)} \leq \cdots \leq p_{(N)}$$

be the ordered $p$-values corresponding to the null hypotheses $H_{0,(1)}, \cdots, H_{0,(N)}$. Rejects the hypotheses $H_{0,(1)}, \cdots, H_{0,(j^*)}$, where $j^*$ is the number such that

$$j^* := \max_{1 \leq j \leq N} \left\{ j : p_{(j)} \leq \gamma_j \right\}.$$

where $\gamma_j := \frac{j}{N} q$ be the rejection criteria. Giglio et al. (2021) suggest modify the method of Benjamini and Hochberg (1995) by precluding the extremely negative alpha funds in advance (in fact fund's $t$ statistics). Define the reduced set of funds indices as

$$\tilde{N} := \bigcup_{1 \leq i \leq N} \left\{ i : t_i > -\log(\log(T))\sqrt{\log N} \right\}$$

and the rejection criteria $\gamma_j$ is therefore change to $\frac{j}{|\tilde{N}|} q$. They show by theoretic inference and Monte Carlo simulation that this alpha screen procedure improves test power while remaining controlling for the FDR under $q$ level.

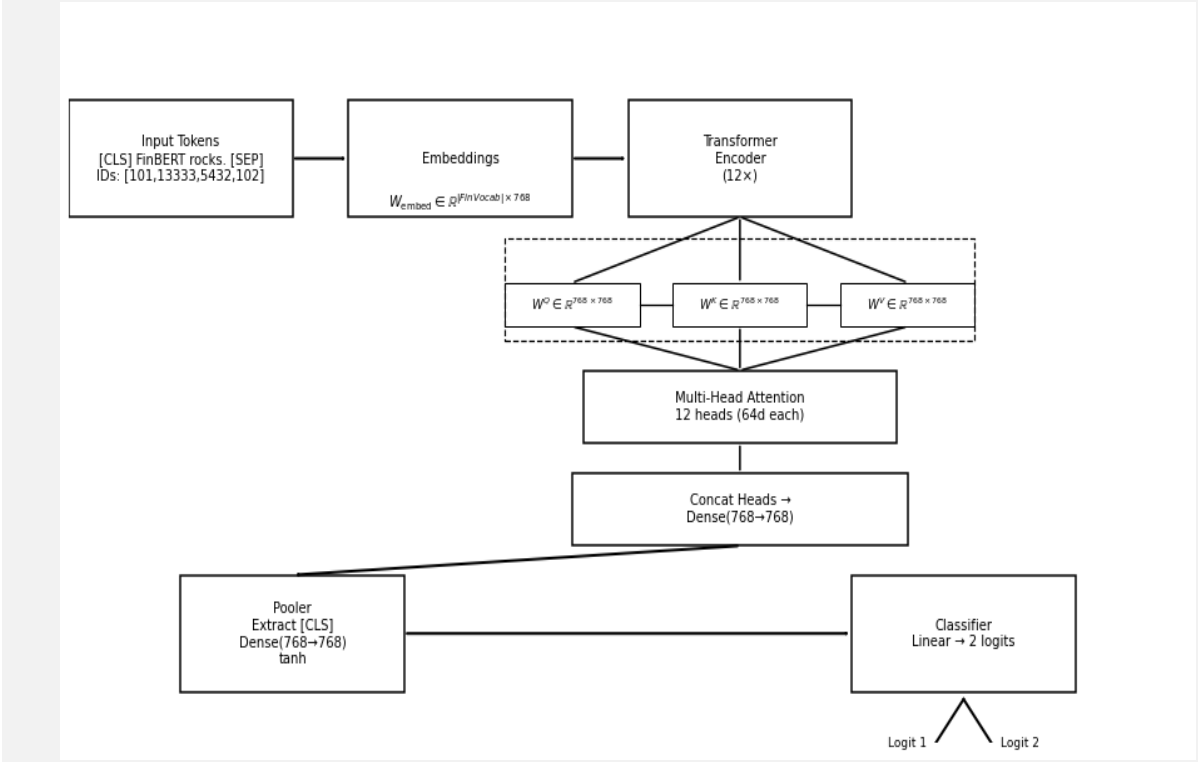
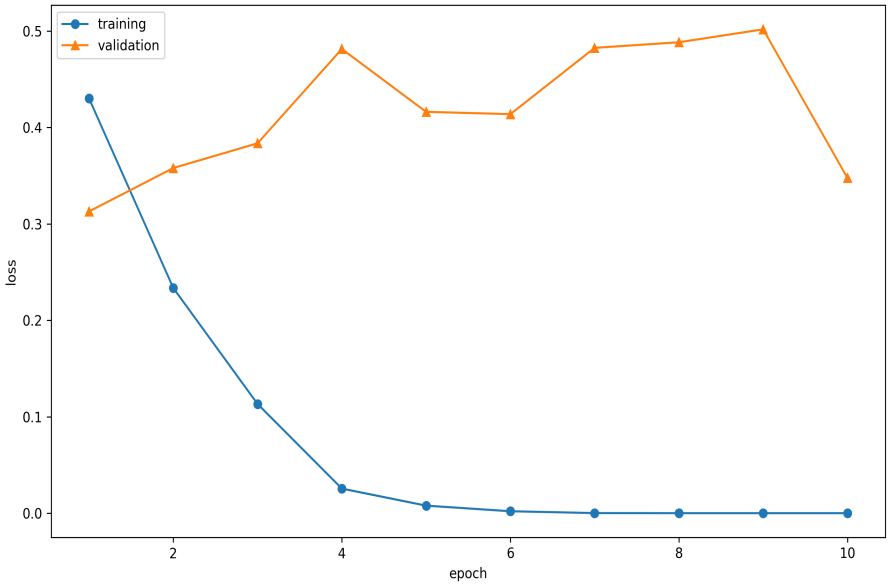
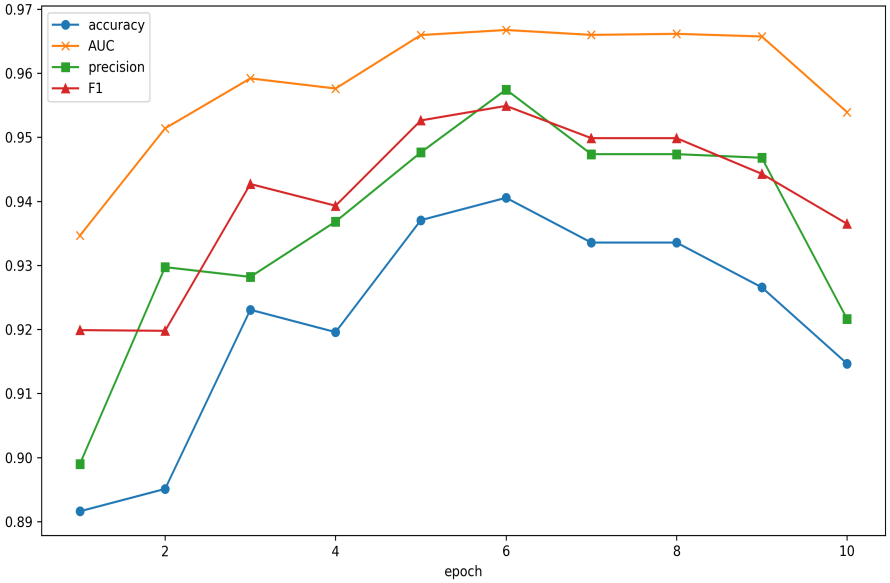



Figure A.1: Architecture of the fine-tuned BERT to binary classification model.




(a) Training and validation loss by epoch




(b) Prediction performance on the validation sample by epoch

Figure A.2: Training progress and validation performance across epochs

Table A.1: Classification performances measures of nested 10-fold cross validation

| | Linear | Logit | LDA | KNN | SVM | RF | Tree | GB |
|---|---|---|---|---|---|---|---|---|
| Panel A: Accuracy | | | | | | | | |
| Median | 0.70 | 0.70 | 0.68 | 0.80 | 0.84 | 0.86 | 0.76 | 0.84 |
| Mean | 0.70 | 0.70 | 0.68 | 0.81 | 0.84 | 0.86 | 0.77 | 0.84 |
| Max | 0.77 | 0.75 | 0.76 | 0.87 | 0.86 | 0.90 | 0.80 | 0.88 |
| Min | 0.65 | 0.65 | 0.65 | 0.79 | 0.80 | 0.82 | 0.73 | 0.81 |
| STD | 0.04 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.02 |
| Panel B: AUC | | | | | | | | |
| Median | 0.70 | 0.70 | 0.66 | 0.73 | 0.80 | 0.82 | 0.73 | 0.81 |
| Mean | 0.70 | 0.70 | 0.66 | 0.74 | 0.80 | 0.82 | 0.74 | 0.80 |
| Max | 0.78 | 0.75 | 0.74 | 0.81 | 0.84 | 0.88 | 0.80 | 0.84 |
| Min | 0.65 | 0.65 | 0.61 | 0.72 | 0.75 | 0.79 | 0.70 | 0.74 |
| STD | 0.04 | 0.03 | 0.04 | 0.03 | 0.03 | 0.02 | 0.03 | 0.03 |
| Panel C: Precision | | | | | | | | |
| Median | 0.82 | 0.82 | 0.77 | 0.79 | 0.85 | 0.86 | 0.81 | 0.84 |
| Mean | 0.82 | 0.82 | 0.77 | 0.79 | 0.85 | 0.86 | 0.82 | 0.85 |
| Max | 0.89 | 0.86 | 0.84 | 0.84 | 0.88 | 0.91 | 0.86 | 0.89 |
| Min | 0.78 | 0.76 | 0.73 | 0.77 | 0.79 | 0.82 | 0.80 | 0.82 |
| STD | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 | 0.02 | 0.02 | 0.03 |
| Panel D: F1 Score | | | | | | | | |
| Median | 0.75 | 0.75 | 0.75 | 0.87 | 0.88 | 0.90 | 0.82 | 0.88 |
| Mean | 0.75 | 0.75 | 0.75 | 0.87 | 0.88 | 0.90 | 0.83 | 0.88 |
| Max | 0.82 | 0.80 | 0.81 | 0.91 | 0.90 | 0.93 | 0.85 | 0.91 |
| Min | 0.70 | 0.71 | 0.71 | 0.86 | 0.86 | 0.87 | 0.79 | 0.86 |
| STD | 0.04 | 0.03 | 0.03 | 0.01 | 0.01 | 0.02 | 0.02 | 0.01 |

This table reports the classification performance measures: Median, Mean, Max, Min, and SD, which are the median, average, maximum, minimum, and standard deviation of the outer 10-folds measures. Statistical learning methods include the following. Linear: linear regression; Logit: logistic regression; LDA: linear discrimination analysis; KNN: $k$-nearest neighbor approach; SVM: support vector machine with the Gaussian kernel; RF: random forest; Tree: classification tree; GB: gradient boosting.

Table A.2: Summary statistics of risk factors: 1994-2015

Panel A: Risk factors

| | N | Mean | STD | Min | Median | Max |
|---|---|---|---|---|---|---|
| MKTRF | 263 | 0.619 | 4.421 | -17.230 | 1.320 | 11.350 |
| SMB | 263 | 0.146 | 3.393 | -17.170 | 0.000 | 22.080 |
| HML | 263 | 0.192 | 3.094 | -11.250 | -0.020 | 12.910 |
| MOM | 263 | 0.489 | 5.161 | -34.580 | 0.580 | 18.380 |
| CS | 263 | -0.053 | 0.208 | -0.804 | -0.060 | 1.532 |
| Δ10Y | 263 | -0.013 | 0.227 | -1.110 | -0.030 | 0.650 |
| PTFSBD | 263 | -1.492 | 15.251 | -26.630 | -3.860 | 68.860 |
| PTFSFX | 263 | -0.700 | 19.492 | -30.130 | -5.180 | 90.270 |
| PTFSCOM | 263 | -0.366 | 14.255 | -24.650 | -3.010 | 64.750 |
| PTFSIR | 263 | -0.891 | 25.708 | -35.130 | -6.040 | 221.920 |
| PTFSSTK | 263 | -4.876 | 14.056 | -30.190 | -6.990 | 66.620 |

Panel B: Correlation matrix

| | MKTRF | SMB | HML | MOM | CS | Δ10Y | TFSBD | PTFSFX | PTFSCOM | PTFSIR | PTFSSTK |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MKTRF | 1.000 | | | | | | | | | | |
| SMB | 0.208 | 1.000 | | | | | | | | | |
| HML | -0.157 | -0.323 | 1.000 | | | | | | | | |
| MOM | -0.271 | 0.104 | -0.199 | 1.000 | | | | | | | |
| CS | -0.140 | -0.066 | -0.056 | 0.043 | 1.000 | | | | | | |
| Δ10Y | 0.101 | 0.101 | -0.039 | -0.072 | 0.629 | 1.000 | | | | | |
| PTFSBD | -0.255 | -0.057 | -0.092 | 0.017 | -0.026 | -0.180 | 1.000 | | | | |
| PTFSFX | -0.206 | -0.001 | -0.005 | 0.120 | 0.047 | -0.177 | 0.298 | 1.000 | | | |
| PTFSCOM | -0.179 | -0.054 | -0.057 | 0.192 | 0.040 | -0.115 | 0.194 | 0.341 | 1.000 | | |
| PTFSIR | -0.269 | -0.098 | 0.006 | -0.004 | 0.141 | -0.170 | 0.204 | 0.249 | 0.228 | 1.000 | |
| PTFSSTK | -0.238 | -0.089 | 0.099 | 0.000 | 0.056 | -0.198 | 0.223 | 0.254 | 0.174 | 0.332 | 1.000 |

This table reports the summary statistics for the risk factors for the sample period from January 1994 to November 2015. SMB, HML, and MOM are the mimicking portfolios for size, book-to-market, and momentum in stock returns, respectively. PTFSBD, PTFSFX, PTFSCOM, PTFSIR, and PTFSSTK are the returns for the long position of the look back straddles of bonds, currencies, commodities, short-term interest rates, and stocks, respectively. CS is the credit risk factor that is defined as the monthly change in the difference between the BAA bond yield and the 10-year constant maturity Treasury yield (GS10). Δ10Y is the difference between the current GS10 and the lag one GS10.