Technische Universität Dresden

Faculty of Business and Economics
Chair of Economics, esp. International Monetary Economics

Modeling Dependencies Between Narratives, Emotions, and Fundamentals in Stock Prices Using Transformers

Master's Thesis

submitted by

Lars Augustat

Student ID: 3096551

M.Sc. Economics, 7th Semester

First Supervisor:	Prof. Dr. Stefan Eichler
Second Supervisor:	Ibrahim Bozdemir
Submission Date:	20.10.2025

Contents

Abbreviation	Full Term
AINI	AI Narrative Index
AIHI	AI Hype Index
AI	Artificial Intelligence
ADF	Augmented Dickey–Fuller Test
AMEX	American Stock Exchange
ANN	Artificial Neural Network
APT	Arbitrage Pricing Theory
BERT	Bidirectional Encoder Representations from Transformers
BH	Benjamini–Hochberg Method
BoW	Bag-of-Words Model
CAPM	Capital Asset Pricing Model
CCAPM	Consumption Capital Asset Pricing Model
CMA	Conservative Minus Aggressive
CMA	Capital Allocation Line
DJIA	Dow Jones Industrial Average
DDM	Dividend Discount Model
EMH	Efficient Market Hypothesis
EMA	Exponential Moving Average
EMAs	Exponential Moving Averages
ETF	Exchange-Traded Fund
FF3	Fama–French Three-Factor Model
FFFM	Fama–French Five-Factor Model
FinBERT	Financial BERT
FDR	False Discovery Rate
FT	Financial Times
FOMO	Fear of Missing Out
GELU	Gaussian Error Linear Unit
GI	General Inquirer
GISI	General Investor Sentiment Index
GPR	Geopolitical Risk Index
GPMS	Google-Profile of Mood States
GPT	Generative Pre-trained Transformer
ICAPM	Intertemporal Capital Asset Pricing Model
IPO	Initial Public Offering
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
LSTM	Long Short-Term Memory
MA	Moving Average
MAG7	Magnificent Seven
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MKT	Market Factor

Abbreviation	Full Term
NLP	Natural Language Processing
NYSE	New York Stock Exchange
NSR	No Significant Results
OVB	Omitted Variable Bias
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PP	Phillips–Perron Test
RMW	Robust Minus Weak
RMSE	Root Mean Squared Error
SDF	Stochastic Discount Factor
SMB	Small Minus Big
SOX	PHLX Semiconductor Index
TF–IDF	Term Frequency–Inverse Document Frequency
VIX	CBOE Volatility Index
WSJ	Wall Street Journal

1 Introduction

Modern financial markets process vast volumes of information and capital daily, serving as central mechanisms for resource allocation, distribution, and economic stabilization. Their efficiency and stability depend not only on institutional frameworks but also on the behavioral and cognitive patterns of market participants. Understanding the psychological and informational foundations of these decisions is therefore of both theoretical and practical importance (Greenwood & Smith, 1997; Shiller, 2021), particularly amid rapid technological transformations such as the rise of Artificial Intelligence (AI) (Floridi, 2024; Widder & Hicks, 2024).[1]

Traditional finance theory, most prominently represented by the Efficient Market Hypothesis (EMH) (Fama, 1970; Malkiel, 2003), assumes that investors act rationally and that prices fully and instantaneously reflect all available information. Behavioral finance challenges this view by highlighting systematic deviations from rationality driven by cognitive biases, emotions, and collective narratives (Hirshleifer, 2015; Shiller, 2020).

This thesis contributes to this debate by empirically examining how emotions and narratives shape asset prices, focusing on the emergence of AI as a highly transformative yet fundamentally uncertain technology (Floridi, 2024; Nathan et al., 2024).

Using advanced Natural Language Processing (NLP)-realted technologies - specifically Transformer-based models - the study constructs four sentiment indices collectively referred to as the AI Narrative Index (AINI). These indices quantify the prevalence and emotional tone of AI-related narratives in the Wall Street Journal (WSJ).

The WSJ is selected because it is (i) among the most influential financial newspapers worldwide (Mai & Pukthuanthong, 2021), (ii) has been widely used in prior literature to infer market sentiment and narratives (e.g. S. Baker et al., 2016; Bybee et al., 2020; Lachana & Schröder, 2025; Tetlock, 2007), and (iii) provides technically accessible large-scale data (see Section 5.1).

Employing daily news content as a proxy for investor sentiment about AI builds on the premise that financial media both reflect and shape market expectations. The reflective aspect arises from the observation that news media are demand-driven and partly intentionally mirror existing beliefs, as readers derive utility from consuming information that confirms their priors (Mullainathan & Shleifer, 2005; Shiller, 2021). The formative aspect has likewise been demonstrated empirically (Engelberg & Parsons, 2011), as discussed in greater detail in Section 4.

To examine whether such narratives hold predictive power for market outcomes, Granger-causality analysis is applied. The method tests if process X improves forecasts of Y beyond information already contained autoregressively via Y's own past, making it well suited to assess the EMH's core claim that all relevant information is already incorporated into prices. To get a better understanding of underlying dynamics, the reverse direction is also tested.

[1] According to the Cambridge Dictionary (n.d.), AI refers to "the use of computer programs that have some of the qualities of the human mind, such as the ability to understand language, recognize pictures, and learn from experience." While this thesis focuses mostly on the perceived impact of AI, this is generally in line with how the technology is understood by the public. A deeper investigation is beyond the scope of this paper.

Guided by insights from behavioral finance (see section 2.4), this thesis is organized around two central hypotheses:

- **H1:** AI-related narratives exert a positive effect on related asset prices on the subsequent trading day, implying slower information incorporation than predicted by the EMH.
- **H2:** As investors update their beliefs, prices exhibit mean-reverting behavior in the days following the initial reaction.

It is important to emphasize that H1 and H2 imply that markets *follow* narratives, i.e., that causality runs from the AINI to returns.

However, here it is important to note that (i) as discussed in Section 3.3, the interplay between narratives and markets is likely more complex, and (ii) this thesis does not assume causality in the strict structural sense but rather in the sense of *Granger causality*,[2] as unobserved confounding variables may still influence both processes.

This thesis makes three main contributions. First, it introduces the AINI, a novel, technology-specific narrative measure that extends beyond the generalized sentiment indices currently used in empirical finance.[3] Second, it applies state-of-the-art NLP methods based on Transformer architectures, which - despite their proven success in other domains - remain underutilized in financial sentiment and narrative research (Acheampong et al., 2021; Mishev et al., 2020). Third, it employs Transformers to explicitly model narratives around a specific technology - according to the author's knowledge, a novelty within the NLP literature itself.[4]

The remainder of this thesis is structured as follows. Section 2.1 introduces the theoretical foundations of market efficiency, followed by Section 2.2 on theories of no-arbitrage. Section 2.4 then discusses approaches that explicitly incorporate inefficiencies, while and behavioral dynamics.

The subsequent sections describe the data collection, the construction of the AINI (section 6.1), and the empirical methodology, including the Transformer-based modeling and Granger-causality framework.

Next Section 4 reviews the empirical literature on narratives and sentiment in financial markets.

Finally, in section 6 and it's subsections, results are presented and discussed in light of both behavioral and no-arbitrage theories, followed by concluding remarks and avenues for future research.

[2] A process A is said to Granger-cause a process B if lagged values of A contain information that helps predict future values of B, beyond what is contained in B's lags. The key distinction from true causality is that Granger causality concerns predictive relationships and does not imply any underlying causal mechanism (Granger, 1969).

[3] To the best of the author's knowledge, no existing study empirically constructs a technology-specific narrative or sentiment index.

[4] While this may sound redundant, it is not trivial: the first novelty refers to constructing a technology-related narrative index relevant for financial markets, whereas the third refers to a Transformer-based narrative measure for any technology-related topic.

2 Theoretical Background: Asset Pricing

This section discusses asset pricing in context of various theories. Subsection 2.1 will discuss the canonical EMH, which has been one of the most (if not the most) influential theory in Finance over the last decades (Sewell, 2011). Next, in subsection 2.2 closely-related theories and models building up on the idea of no-arbitrage will be examined.[5] Having thus evaluated arguments in favor of zero-arbitrage, subsection 2.3 will discuss a departure from those theories, which has gained some momentum since the early 1990s, i.e. theories assuming financial markets can significantly diverge from inefficiency. Building up on this, subsection 2.4 will investigate further evidence from the behavioral science which might indicate psychological and sociological departures from efficiency. As we will see, emotions and narratives play a key role in theories and models indicating inefficiencies.

2.1 Efficient Market Hypothesis

The EMH has long served as the central paradigm in financial economics (Malkiel, 2003). It posits that, in competitive financial markets, asset prices fully and instantaneously reflect all available information.[6].

In formal notation, let I_t denote the information set available at time t, then the price of an asset with payoff X_{t+1} satisfies:

$$P_t = \mathbb{E}[M_{t+1} X_{t+1} \mid I_t] \tag{1}$$

where, M_{t+1} denotes the Dividend Discount Model (DDM), or pricing kernel, which captures how investors discount future payoffs conditional on information available at t. It reflects both the time value of money (through inter-temporal substitution) and compensation for bearing risk (Cochrane, 2000, p. 9).

From a practical perspective, EMH postulates that investors behave *asymptotically* rational and price in new information efficiently by buying and selling (Fama, 1970). While some investors will overestimate a given stock, others will underestimate it, but on average the estimate will be the best approximation of its fundamental value, which is sometimes also called *Fair Value* (Malkiel, 2003).

In this framework, deviations from Fair Value are considered temporary and self-correcting. Arbitrage is expected to eliminate mispricings before they can persist. Importantly, proponents of the EMH do not claim that markets are perfectly rational in every moment. Rather, they argue that markets behave *as if* investors are rational, because arbitrage and competition drive prices toward a state that approximates informational efficiency (Fama, 1970, 1991; Malkiel, 1989).

Furthermore Fama (1970) distinguished between three forms of efficiency, depending on the information set I_t:

[5] We can think of EMH as an arbitrage-free theory with stronger assumptions than purely no-arbitrage theories and models.

[6] Here, "instantaneously" should be understood in a practical sense: price adjustments occur quickly enough that even professional investors cannot systematically earn arbitrage profits. While this assumption may sound strong, there is ample evidence that even professional investors do not consistently outperform the market. See below and Malkiel (2019)

- **Weak form:** I_t includes only past prices and returns. This implies that technical analysis alone cannot generate excess returns.
- **Semi-strong form:** I_t includes all publicly available information, such as firm-specific reports, analyst forecasts, and other relevant news. This implies that even fundamental analysis cannot systematically yield above-market returns.
- **Strong form:** I_t contains both public and private information. This would imply that even insider trading cannot yield persistent abnormal profits.

Importantly, none of the forms of efficiency can be tested independently of an asset pricing model. As Fama (1970, 1991) emphasized, all tests of efficiency are necessarily joint tests of (i) informational efficiency and (ii) a specific model of M_t. If a return pattern appears predictable, it may reflect mispricing (inefficiency), but it could equally reflect a missing risk factor or a wrongly defined DDM. This is the so-called *joint-hypothesis problem*, which is central to modern empirical finance (ibid.).

To illustrate the problem, it is instructive to start with the canonical DDM, in which the price P_t of a stock at time t equals the expected present value of all future dividends D_{t+k}, discounted at the constant rate of return r required by investors:

$$P_t \;=\; E_t\left[\sum_{k=0}^{\infty} \frac{D_{t+k}}{(1+r)^{k+1}}\right] \tag{2}$$

Here, r represents the investors' required rate of return, capturing both the time value of money and compensation for risk. Under this formulation, the value of a stock is entirely determined by expectations of future dividends and the discount rate applied to them (Williams, 1938).

To sharpen intuition, we follow Campbell and Shiller (1988) and Shiller (1981) in log-linearizing the price-dividend identity and decomposing it into expected changes in future payoffs and expected changes in future discount rates. Using a log-linear approximation (see Cochrane, 2000, 154 ff.), the identity can be expressed as:

$$(p_t - d_t); \approx; k + (1-\rho)\sum_{j=1}^{\infty} \rho^{j-1} E_t\left[\Delta d_{t+j}\right] - (1-\rho)\sum_{j=1}^{\infty} \rho^{j-1} E_t\left[r_{t+j}\right] \tag{3}$$

Here, p_t denotes the log stock price, d_t the log dividend, ρ is a linearization parameter close to but less than one, and r_{t+j} is the log return. The term $\Delta d_{t+j} \equiv d_{t+j} - d_{t+j-1}$ represents the change in log dividends, i.e. the dividend growth rate.

This decomposition highlights that the value of a stock is fundamentally driven by two distinct components: (i) expectations of future dividend growth (Δd_{t+j}), and (ii) expectations of future discount rates (r_{t+j}). While dividends represent the periodic cash flows paid to shareholders, the expected discount rates are critical, since they reflect the premium investors demand for deferring consumption and bearing risk. In terms of Equation 1, they correspond to the properties of the DDM, since expected returns are determined by the covariance between asset payoffs and the stochastic discount factor (see below).

What, then, determines the DDM? A broad class of asset pricing models can be viewed as attempts to characterize or approximate it. Under the EMH, they all collapse to a fundamental relationship:

$$P_t \;=\; \mathbb{E}_t\left[\beta\,\frac{u'(c_{t+1})}{u'(c_t)}\,X_{t+1}\right] \tag{4}$$

Here, β denotes the subjective discount factor, $u'(c_t)$ the marginal utility of consumption at time t, and X_{t+1} again the asset payoff. Rearranging yields a classical definition of the DDM:

$$m_{t+1} \;=\; \beta\,\frac{u'(c_{t+1})}{u'(c_t)} \tag{5}$$

i.e the representation of the DDM as the marginal utility of consumption in $t+1$ relative to that of t.

Expressed more intuitively, the condition is almost tautological: the utility lost from reducing consumption today must equal the utility gained from additional consumption tomorrow.[7]

This condition - often referred to as the *first-order condition for optimal consumption and portfolio choice* - is, as Cochrane (2000, p. 15) emphasizes, the cornerstone of modern asset pricing. In his words, "[m]ost of the theory of asset pricing just consists of specializations and manipulations of this formula."

With the stochastic discount factor established as the general foundation, the next step is to examine how the EMH - and, more broadly, no-arbitrage models - translate this into concrete asset pricing frameworks.

2.2 Asset pricing without arbitrage

A classical point of departure is Markowitz (1952): investors have concave preferences and choose portfolios to balance expected return against risk. Under quadratic utility or (approximately) elliptical return distributions, risk can be measured by the variance of returns, so investors solve a mean–variance problem: minimize variance for a given expected return, or equivalently maximize expected return for a given variance.[8]

Since asset returns are not perfectly correlated, diversification lowers risk for any given level of return, leading investors to hold broad portfolios. This principle underlies what is commonly referred to as *Modern Portfolio Theory* (Elton & Gruber, 1997).

Graphically, the set of all feasible combinations of risky assets forms the *mean–ier*, i.e. the hyperbola in figure 1. The *efficient frontier* is its upper part, consisting of portfolios that, for each level of risk, deliver the maximum attainable expected return (Markowitz, 1952).

[7] While individual investors naturally aim to maximize personal payoffs, under market efficiency the equilibrium condition (5) must hold.

[8] Extensions consider higher moments of the return distribution, e.g. skewness preferences in Kraus and Litzenberger (1976).

Introducing a risk-free asset improves this trade-off: investors can now achieve a superior risk–return combination by holding the risk-free asset together with a portfolio of risky assets. In this setting, all efficient portfolios lie on the Capital Allocation Line (CAL), the straight line through the risk-free rate that is tangent to the efficient frontier.

Crucially, the tangency point corresponds to a unique risky portfolio that maximizes the risk-return trade-off (i.e., the Sharpe Ratio, see below). All investors hold this same risky portfolio, differing only in the weights they assign to it and to the risk-free asset, depending on their risk appetite (Cochrane, 2000, p.80 ff.). An easy, intuitive geometrical interpretation is given below, where the Y-Axis depicts the expected return and the X-Axis the risk, measured in standard deviations.

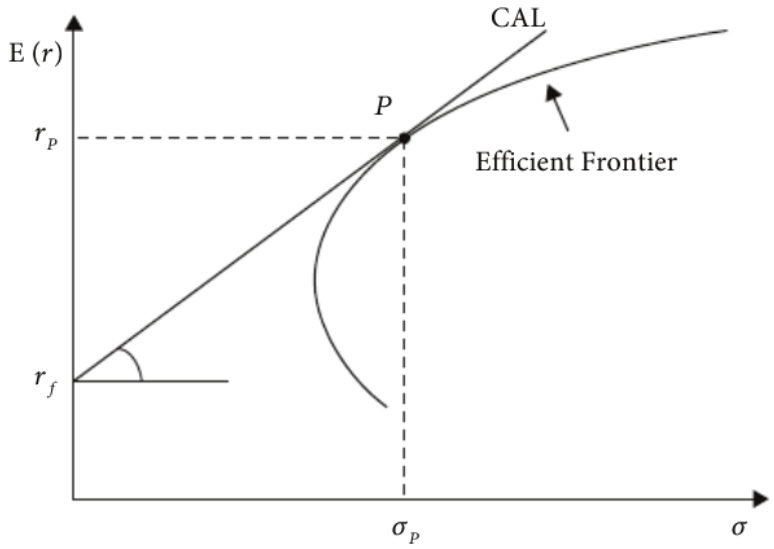

Figure 1. Mean–variance frontier. *Source:* Ahmadi and Peivandizadeh (2022).

In equilibrium, under homogeneous expectations and market clearing, this tangency portfolio is equal to the *market portfolio*, providing the bridge to the Capital Asset Pricing Model (CAPM).

The CAPM builds on these mean–variance foundations and shows that in equilibrium, expected excess returns, i.e. returns above the risk-free rate, are proportional to an asset's covariance with the market portfolio .[9] Formally:

$$E[R_i] - R_f = \beta_i \left(E[R_m] - R_f\right) \tag{6}$$

where R_i denotes the return of portfolio i, R_f the risk-free rate, R_m the return on the market portfolio,[10] and $\beta_i = \frac{\text{Cov}(R_i, R_m)}{\text{Var}(R_m)}$ measures the exposure of asset i to systematic risk. The slope of the CAL is the so called Sharpe ratio of the market portfolio, $\frac{E[R_m]-R_f}{\sigma_m}$, i.e. the excess return per unit of risk (as measured by variance) taken.

[9] The model assumes a world of homogeneous expectations, two periods, perfect competition, and zero transaction costs.

[10] The problem of defining the market portfolio is non-trivial. A key issue is to define such a portfolio *ex ante*, without ex post cherry-picking. For discussion see (Cochrane, 2000, p. 120ff.).

Simply speaking, because stock-specific, idiosyncratic risk can be diversified away, an investor will be rewarded by excess returns when holding excess *systematic risk*. Here systematic risk is measured as the stock's co-movement with the market. An asset with $\beta = 1$ delivers the same expected excess return as the market - as it moves in synchrony with the market. An asset with $\beta = 2$ tends to move twice as strongly as the market and is rewarded with twice the market's expected excess returnm (Lintner, 1975; Mossin, 1966; Sharpe, 1964).

Although the CAPM fails to explain the full cross-section of asset returns[11], it has been highly influential as a benchmark model. It provides both an intellectual framework for understanding risk-return trade-offs and a starting point for empirical investigations into anomalies unexplained by β (Fama, 1970, 1991; Malkiel, 2003).

A key limitation of the CAPM is its static, two-period structure. It cannot capture the fact that investors care not only about next-period returns, but also about how today's portfolio choices affect a large set of possible futures. In particular, the CAPM does not allow for hedging against long-run risks.

Extending the model to a dynamic, multi-period setting, Merton (1973) developed the Intertemporal Capital Asset Pricing Model (ICAPM). In this framework, investors maximize expected lifetime utility, taking into account that investment opportunities may change over time. As a result, asset prices are determined not only by covariances with the market portfolio, but also by covariances with *state variables* that describe the evolution of investment opportunities (e.g. interest rates, volatility, labor income, or technological shocks).

Formally, the inter-temporal stochastic discount factor can be written as

$$m_{t+1} = \beta \frac{u'(g(z_{t+1}))}{u'(g(z_t))} \tag{7}$$

where $u'(\cdot)$ denotes marginal utility, and $g(z_t)$ is a function mapping the state variables z_t to consumption opportunities. This formulation generalizes the consumption-based DDM in (**??**) by recognizing that changes in state variables affect the marginal utility of wealth across time. In equilibrium, expected excess returns are thus affected not only by their covariance with the market, but also with these additional state variables (Cochrane, 2000, p.165).

Besides ICAPM, another canonical model is the Consumption Capital Asset Pricing Model (CCAPM), which extends the insights of the CAPM to an explicitly inter-temporal setting of consumption optimization (Cochrane, 2000, p.22 ff.). While thereby neglecting other state variables (albeit one might argue that those are, in the end, related back to an investor's consumption), the CCAPM identifies the (stochastic) marginal utility of consumption as the fundamental driver of asset prices. Building on the above-mentioned first-order condition formalized in equation (5)), the model explicitly links $E[R_i] - R_f$ to the covariance of asset payoffs with aggregate consumption growth.

As in the ICAPM, the key implication is that not all co-movements with the market are judged as being equal: assets delivering high payoffs precisely in "bad times" (such as recessions) - when marginal utility of consumption is high - are especially valuable, since

[11] The cross-section of return refers to the difference in returns between different stocks.

they provide a hedge against those bad times. Conversely, assets that tend to pay off in "good times" are less desirable and must offer higher excess returns. Formally, the model can be written as:

$$E[R_{i,t+1} - R_{f,t+1}] = \gamma \operatorname{Cov}(\Delta c_{t+1}, R_{i,t+1}) \tag{8}$$

where Δc_{t+1} denotes aggregated consumption growth, γ models the investor's relative risk aversion, and $R_{i,t+1}$ denotes the return on asset i (Breeden et al., 1989).

While the concept of the CCAPM is theoretically interesting, its empirical explanatory power has been rather limited. Or, as Cochrane (2000, p.47) concludes "[t]he consumption-based model is, in principle, a complete answer to all asset pricing questions, but works poorly in practice. This observation motivates other asset pricing models."

To solve the problem of limited explanatory power, many models have been developed. One of the most influential is the Fama–French Three-Factor Model (FFTM) (Fama & French, 1993), which is discussed next.

The model originates from two empirical observation of persistent anomalies in stock returns: (i) firms with small market capitalization (small-cap stocks) tended to outperform large-cap firms, and (ii) firms with high book-to-market ratios (value stocks) tended to outperform those with low book-to-market ratios (growth stocks). These empirical regularities could not be explained within the framework of the CAPM.[12].

Importantly, FFTM is not derived from utility theory or first principles, but rather as an empirical adaptation based on the above mentioned anomalies. It therefore belongs to the class of relative asset pricing models: it seeks to explain expected returns across assets through covariances with empirically chosen risk factors, rather than deriving absolute price levels. While framed as an extension of the CAPM, the FFTM is conceptually closer to the Arbitrage Pricing Theory (APT) (Cochrane, 2000, p.170-172), which will be discussed in more detail below.

Formally, the model augments the CAPM's *beta* by introducing two additional factors: the return differential between diversified portfolios of small- and large-cap firms (Small Minus Big (SMB)) and the return differential between high and low book-to-market firms (High Minus Low (HML)). The construction of these factors proceeds as follows.

At the end of each June, all New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and National Association of Securities Dealers Automated Quotations (NASDAQ) common stocks are independently sorted on size and book-to-market equity.[13] Size is defined by market capitalization, with the median NYSE firm as the cutoff between Small (S) and Big (B). Book-to-market is defined by the ratio of book equity to market equity, with the 30th and 70th percentiles of NYSE firms as breakpoints for Low (L), Medium (M), and High (H).

The intersection of these two sorts produces six value-weighted portfolios: S/L, S/M, S/H, B/L, B/M, and B/H, where the first letter indicates firm size and the second letter

[12]Interestingly, the authors start by observing anomalies and then deduce their model. This is, as we will see below, an interesting parallel to the behavioral finance literature (Hirshleifer, 2015, see).

[13]The NYSE, the AMEX, and the NASDAQ are the three major U.S. stock exchanges. Breakpoints for size and book-to-market sorts are based only on NYSE firms, in order to avoid distortions from the large number of very small firms listed on AMEX and NASDAQ.

indicates the book-to-market category. The monthly returns of these portfolios from July of year t through June of $t+1$ are then used to compute the differences in returns. The size premium SMB is defined as the average return on the three small-stock portfolios minus the average return on the three big-stock portfolios, while the value premium HML is the difference between the average returns of the high and low book-to-market portfolios, averaged across size groups.

The expected excess return of portfolio (or asset) i is then modeled as

$$E[R_i] - R_f = \beta_{i,m}\,(E[R_m] - R_f) + \beta_{i,s}\,SMB + \beta_{i,h}\,HML \tag{9}$$

where $\beta_{i,m}$ captures exposure to the market portfolio, $\beta_{i,s}$ to the size premium, and $\beta_{i,h}$ to the value premium. The model is typically tested in time-series regressions of portfolio returns on the factors, with the intercepts (α_i) serving as indicators of mispricings.[14]. Importantly, this implies that the expected return on stock of some company i will *not* depend on it's size, but on it's β on the size factor. If a big companie's stock behaves *like* that of small companies, it will have high loadings (a high beta) on SMB (Cochrane, 2000, p.78-79).[15]

Empirically, the inclusion of SMB and HML improves the model's ability to explain the cross-section of stock returns. From a theoretical perspective, the additional factors are commonly interpreted as proxies for systematic sources of risk not captured by the β of CAPM.

Specifically, SMB is often associated with the heightened sensitivity of smaller firms to economic fluctuations, while HML is rationalized by the idea that value stocks—typically firms in mature or distressed industries—are riskier because they have lower expected growth prospects. In this interpretation, the anomalies of size and value premia do not violate efficiency, but instead represent compensation for systematic risk.

While the FFTM led to important empirical insights, subsequent evidence also revealed contradictions. A notable example is provided by Novy-Marx (2013), who extended the cross-sectional analysis of stock returns to include measures of profitability. He demonstrated that highly profitable firms, despite generally being classified as growth stocks (thus loading low on HML), systematically outperformed unprofitable firms.

Interestingly, this relationship can also be derived from a decomposition of the current market value P_t.[16] Using "clean surplus accounting" (ibid., p. 491), Fama and French (2006) show that the HML (see above) can equivalently be expressed in terms of book equity as

$$\frac{P_t}{B_t} = \sum_{\tau=1}^{\infty} \frac{E(Y_{t+\tau} - \Delta B_{t+\tau})}{(1+r)^{\tau} B_t} \tag{10}$$

[14] In theory, a perfect asset pricing model would yield intercepts of zero, i.e. it would fully explain the variation in expected returns. In practice, however, financial markets are complex and any model is necessarily a simplification of reality. Consequently, the empirical asset pricing focuses on minimizing the pricing errors α_i (Fama & French, 2015, see)

[15] While sounding trivial, this is an important assumption: Investors watch the stocks *true* behavior, not (self-proclaimed) labels. From a different angle, an investor would thus watch if a firm is *truly* a company focused on building AI, and not just a self-proclaimed one.

[16] Note that Fama and French (2006) use M to denote the current market value. To avoid confusion, this thesis adopts P_t for market value and reserves M for the DDM as in Equation 1.

where $Y_{t+\tau}$ denotes expected earnings in period $t+\tau$, $\Delta B_{t+\tau} = B_{t+\tau} - B_{t+\tau-1}$ is the change in book equity, and r is the required rate of return (a deterministic analogue of the DDM in equation 1). This formulation links valuation not only to expected future dividends, but also to profitability (through earnings) and investment (through changes in book equity). The latter reflects the fact that increases in book equity stem either from retained earnings used for investment or from reductions due to depreciation.

Novy-Marx (2013) argues that the profitability premium may indeed represent a risk premium, since more firms with relatively high productivity could exhibit more sensitivity towards fluctuations in the business cycle. At the same time he remains cautious, noting that the pattern could also reflect behavioral (i.e. irrational) aspects. Importantly, the profitability premium challenges the interpretation of the value premium: if high book-to-market firms earn higher returns because they are low-growth, the outperformance of high-profitability growth firms casts doubt on this risk-based mechanism behind HML.

Building on this insight, Fama and French (2015) introduced the Fama–French Five-Factor Model (FFFM), adding proxies for profitability and investment to the explanatory variables. The model takes the following form in time-series regressions:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_{i,m}(R_{m,t} - R_{f,t}) + s_i\, SMB_t + h_i\, HML_t + r_i\, RMW_t + c_i\, CMA_t + \varepsilon_{i,t} \quad (11)$$

where $R_{i,t}$ denotes the return of asset i at time t, $R_{f,t}$ the risk-free rate, $R_{m,t}$ the market return, and $\varepsilon_{i,t}$ the residual. The coefficients $\beta_{i,m}$, s_i, h_i, r_i, and c_i measure the sensitivities of asset i to the market factor (Market Factor (MKT)), size (SMB), value (HML), profitability (Robust Minus Weak (RMW), Robust Minus Weak), and investment (Conservative Minus Aggressive (CMA), Conservative Minus Aggressive), respectively.

As in FFTM, these factors are measured as return differentials between diversified portfolios sorted on the relevant characteristics.[17]

As in the FFTM, the portfolios are formed by sorting firms independently on size and a second characteristic, i.e. by performing separate 2×3 sorts, in which firms are split into S and B based on the NYSE median market capitalization, and independently into three groups (Low, Medium, High) according to profitability or investment. Here SMB is defined as the average of three separate size premiums, one each from the size-B/M, size-profitability, and size-investment sorts.[18] These sorts are repeated each June, and factor return differentials are computed from July of year $t-2$ and $t-1$. The resulting factors are then used in the time-series regressions.

Empirically, the FFFM substantially improves explanatory power relative to the FFTM, reducing the average pricing errors across a large set of portfolios. However, it also raises new questions. In particular, the model performs poorly for investment-heavy micro-cap firms: portfolios of the smallest companies (lowest size quintile) that simultaneously

[17] The first three factors (MKT, SMB, HML) are constructed as described above. Profitability is measured as operating profitability, defined as annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses, all divided by book equity at the end of the prior fiscal year $t-1$ (Fama & French, 2015, p. 5). Investment is measured as *asset growth*, i.e. the change in total assets between fiscal years $t-2$ and $t-1$, scaled by total assets at $t-2$.

[18] The authors also consider intersections based on 2×2×2×2 sorts, discussion of these alternative constructions is beyond the scope of this thesis.

show the highest investment rates tend to earn significantly lower returns than predicted. Moreover, the inclusion of profitability and investment renders HML largely redundant, challenging its interpretation as a fundamental risk proxy.

Theoretically, Fama-French position their factor-models as mimicking the market portfolio and - albeit with some caution - to the ICAPM, where they see their factors as potential proxies for the state variables (Fama & French, 2015). However, one might also argue that neither the FFTM nor the FFFM assume any kind of information efficiency and are just descriptive models closely related to APT, which we will now very briefly explore before discussing behavioral theories.

First of all it is important to note that APT does - in contrast to CAPM, CCAPM and ICAPM - not start with strong assumptions about investor behavior and information efficiency. Strictly speaking - when defining the EMH by this information efficiency as discussed in the last paragraph - APT does not belong the EMH-paradigm. The key assumption behind APT is simply that by the law of one price, arbitrageurs will price away over -or undervalued stocks thereby making above-market returns impossible.[19].

The key idea is that one does not start with restrictive assumptions and derive optimal prices, but with the basic observation that there is significant correlation between stock-prices - especially those with a shared characteristic, such as stocks of software companies. Furthermore, as those prices are also not perfectly correlated, an investor will combine a portfolio such that idiosyncratic stock risks are negligible and only systematic risks persist. Crucially, however, the APT does not assume that those factors are some state variables one hedges against - e.g. unemployment - but that those risk factors simply exists due to co-movement and that all other factors are priced away by arbitrageurs (Dybvig & Ross, 1989; Ross, 2013).

Contrasting this with what we discussed above, it is still evident that APT is closely related to the ICAPM. The main difference is that ICAPM starts with theoretical assumptions based on CAPM, while APT is build up on empirical observations and in principally agnostic to the true underlying mechanisms. From a theoretical point of view Cochrane (2000, p.170) even concludes that the strongest argument in favor for APT to hold is that it is "just another specialization of the consumption based model."

In sum, we can conclude that all models belonging to the EMH-paradigm can be related to (1), i.e. to the idea of the fair value of a stock being given by it's expected future dividends and some sort of compensation for bearing systematic risk. While the models differ in how this risk is defined - market covariance, consumption growth, or other unspecified risk factors - they all implicitly (or even explicitly) assume that arbitrageurs will sport opportunities in mispriced assets and correct them.

Furthermore, FFTM, FFFM and APT - which are closely related to EMH but do not need it to hold completely - also posits that beyond the reward for taking systematic risks, investors won't be able to systematically outperform the market.

While frameworks such as the EMH or APT emphasize pricing under no-arbitrage conditions, they leave little room for systematic, exploitable mispricings caused by investor sentiment. A first step beyond these paradigms is provided by models of noise traders and subsequent models incorporating well-known biases.

[19] As APT is also not about behavioral anomalies let alone psychological insights, it will be discussed in this paragraph. Sorting by "school of thought" one might separate it as a third branch of finance. However, a discussion of this is beyond the scope of this paper.

2.3 Asset Pricing beyond efficiency

An important point of departure from the arbitrage-free models is the literature on noise traders. The unifying feature of these models is the assumption that some investors *systematically* misinterpret or overreact to information, thereby generating demand shocks. The key question then becomes how rational investors interact with noise traders, and how this interaction can sustain persistent mispricing.[20]

An early contribution was made by Black (1986), who introduced the concept of noise, defined as a signal carrying no information - for example, a news headline that containing no genuinely new content. While such events could occur even under the EMH, Black's insight was twofold: (i) distinguishing information from noise is often difficult, and (ii) some investors cannot distinguish at all, and end up trading on noise while believing they act on information.

While informed investors may arbitrage against noise traders, uncertainty about the timing and magnitude of uninformed demand implies that mispricings can persist. Moreover, if the behavior of noise traders is correlated, their collective impact can drive prices far from fundamentals.

Building on this idea, De Long et al. (1990) formally developed what has been coined the Noise Trader approach. In their two-period overlapping generations model, rational investors and noise traders coexist. Rational investors allocate wealth in period t to maximize consumption when they retire at $t+1$, selling their holdings to the next generation of investors. While having the same goals and time horizon, in contrast, Noise traders follow stochastic beliefs with variance σ.

The key insight is that rational investors cannot simply arbitrage away mispricings because of noise trader risk: the possibility that noise traders push prices even further from fundamentals before convergence occurs. This creates an additional source of systematic risk, which rational investors must account for. As a result, expected returns depend not only on some form of systemic risk (as in CAPM or ICAPM) but also on exposure to noise trader risk.[21]

This is further emphasized by Shleifer and Summers (1990), who showed that common sentiment and so-called feedback trading - buying stocks with positive momentum and selling those with negative momentum - can amplify these effects and may even generate speculative bubbles.[22]

From a practical perspective, noise trader risk helped to explain two prominent empirical puzzles: (i) stock prices appear excessively volatile relative to changes in dividends or discount rates (Black, 1986; Campbell & Shiller, 1988; Malkiel, 2003) (see equation (**??**).), and (ii) closed-end funds often trade at persistent discounts or premia relative to the value

[20] The key term is systematically. EMH as well as APT allow for irrational investors. They just postulate that their effect on prices is negligible, as arbitrageurs will correct prices.

[21] The crucial departure from classical EMH assumptions (e.g. Fama, 1970, 1991) is that the behavior of irrational investors is correlated, rather than offsetting one another in the aggregate.

[22] The debate over whether stock market bubbles truly exist lies beyond the scope of this paper. In the context of Shleifer and Summers (1990), bubbles are understood as situations in which stock prices deviate substantially from their fundamental value, typically being far above it. For a broader discussion, see Kindleberger and Aliber (2005).

of their underlying assets.[23]

Of particular importance for subsequent analysis is the work of Shiller (2014). Building on the distinction between rational investors - sometimes referred to as Smart Money - and retail investors,[24] Shiller argued that the classical DDM (equation 2) overlooks what Keynes famously called Animal Spirits,[25] which he defines as "the demand for stocks per share at time t by everyone who is not smart money," emphasizing that such investors constitute "certainly the majority of investors" (Shiller, 2014, p. 474).

Incorporating this idea, the price P at time t can be expressed as

$$P_t = E_t \left[\sum_{k=0}^{\infty} \frac{D_{t+k} + \phi\, A_{t+k}}{(1 + r + \phi)^{\,k+1}} \right] \tag{12}$$

where D_t denotes dividends at time t, A_t captures non-rational demand, and $\phi > 0$ measures the strength of this channel. Compared to the classical DDM, prices no longer reflect only rationally discounted expected dividends but also a speculative components arising from shifts in investor mood.

Why investors might act irrationally in the first place has been addressed by behavioral literature. A prominent theme is *overconfidence*: people tend to overestimate their abilities and underestimate their shortcomings. Furthermore overconfidence is often accompanied by *self-attribution bias*: people attribute successes to their own skill but blame failures on circumstances (Daniel et al., 1998; Hirshleifer, 2015). Empirically, Barber and Odean (2001) showed that men, being on average more likely to show overconfidence, trade more frequently and achieve lower net returns than women.

Building on these insights, Daniel et al. (1998) developed a formal model where investors receive private signals, which they overweight relative to public information. Overconfidence causes them to exaggerate the significance of the private signal, while self-attribution bias leads them to overreact to confirming news and underweight contradictory news. Since true positive and negative signals are evenly distributed, this mechanism can generate short-term momentum followed by long-term reversal in returns.

Another well-documented phenomenon is *conservatism*: people attach too much weight to their prior believes relative to new information, making them stick to too long to existing opinions compared to what a fully-rational Bayesian agent would do (Edwards, 1968). Combined with the *representativeness heuristic* (Tversky & Kahneman, 1974) - the tendency to judge probabilities based on perceived similarity, to infer meaning from

[23] Closed-end funds are funds which issue a fixed number of shares at inception and do not redeem them before termination, so investors can only trade their shares on secondary markets. Since they regularly report their net asset value, often on a weekly or even daily basis (Pontiff, 1995), comparisons between market prices and fundamental value are straightforward. This was explicitly analyzed by C. M. Lee et al. (1991), who argued that closed-end funds often trade below their fundamental value and exhibit substantially higher price volatility than their underlying assets because they are disproportionately held by retail investors. However, note that there are also EMH-compatible explanations, such as liquidity (Cherkes et al., 2008) and tax effects (Malkiel, 1977).

[24] Shiller avoids the term Noise Traders, reserving it for the specific framework of Shleifer and Summers (1990)

[25] Keynes introduced Animal Spirits to capture the role of sentiment and non-rational motives in economic decision-making.

superficial class assignments, to see patterns in noise and to neglect base rates - this produces predictable misjudgments.

Barberis et al. (1998) incorporated these biases into a formal model of a representative investor. In the model, the agent under-reacts to isolated news (conservatism), but after a sequence of similar signals, overreacts due to representativeness. Simulations show that this framework can reproduce both short-term momentum and long-term reversals.

In a related model, Hong and Stein (1999) introduced two types of traders: news-watchers, who under-react to new information, and momentum traders, who exploit apparent trends without distinguishing whether they stem from fundamentals or noise. Their interaction generates gradual information diffusion, trend-chasing, and eventual reversals.[26]

In sum, the noise trader framework and Shiller's extension of the DDM illustrate how irrationality can enter asset prices alongside fundamentals. Bias-based models such as those of Barberis et al. (1998) and Hong and Stein (1999) further show how systematic psychological tendencies can explain anomalies like momentum and reversals. Taken together, these approaches highlight that prices need not solely reflect discounted fundamentals, but also sentiment, biases, and feedback dynamics.

Beyond those formal models, there exists a broader literature documenting how sentiment and narratives might directly influence - at least some - investors.

2.4 Emotions, Beliefs, and Narratives in Asset Pricing

This section surveys central insights from behavioral theories. While those are not specific to financial markets, their implications for asset pricing can be interpreted as determinants of A_t in equation (12).

However, as we proceed to examine the measurement of investor sentiment and narratives, it is important to acknowledge that especially the term *sentiment* lags an uniformly accepted definition in economics. In some studies (Zhou, 2018, e.g.), it is defined, following Shleifer and Summers (1990), as the deviation between a noise trader's (or more broadly, an irrational investor's) estimate of an asset's fair value and the fair value as assessed by rational investors . Beyond this theoretical framing, especially in empirical work, investor sentiment is often understood more loosely as a general positive or negative attitude toward a particular stock or class of stocks. This is commonly operationalized through proxies that capture investor "bullishness" or "bearishness" (e.g., Zhuge et al., 2017). In other cases, the term is used to refer more directly to the emotional states of investors (Brown & Cliff, 2004).

The concept of narrative is similarly underdefined - but in this case it is so because the general literature on narratives in economics is very limited. Shiller (2020), one of the most prominent proponents of narrative economics, builds up on the Oxford English Dictionary, which defines a narrative as "a story or representation used to give an explanatory or justification account of a society, period, etc." (Oxford English Dictionary, quoted after (Shiller, 2020, p.xvii)) to which he adds that such "stories are not limited to simple chronologies of human events. A story may also be a song, joke, theory, explanation, or plan that has emotional resonance and that can easily be conveyed in casual

[26]Interestingly, the momentum traders start by correcting the under-reaction - thus mimic rational behavior at first - but then start chasing the trend of other momentum traders.

conversation" (ibid.) This understanding is broadly consistent with how other authors have used the concept (Taffler et al., 2024; Tuckett, 2011).

The definition of *sentiment* employed in this thesis departs slightly from that of Shleifer and Summers (1990). Here, even rational yet optimistic (or pessimistic) judgments are classified as *positive* (or *negative*) rather than *neutral* sentiment. As will be discussed in Section **??**, this adjustment is necessitated by the measurement approach: FinBERT evaluates the polarity of textual expressions - how positive or negative they are - without assessing the truth of the underlying statement.

Moreover, many statements occupy a conceptual gray zone and cannot be judged *ex ante* as either rational or irrational. Consider, for instance, the sentence: "NVIDIA will grow exceptionally, providing the chance to yield returns of above 300%." From a purely sentimental perspective, the tone is clearly positive. But is it rational? Relative to market expectations, such growth appears highly improbable. Yet, given the company's realized performance over the past ≈ 2.5 years (see Table **??**), this projection would, in retrospect, have been an *underestimation*.

For *narratives*, the definition by Shiller (2020) is adopted.[27]

Now turning our attention to the behavioral literature, according to Kahneman (2011) dual-process theory, our cognition operates through two systems: the fast, intuitive, emotionally driven System 1, and the slower, analytical, reflective System 2. In emotionally charged or ambiguous situations, System 1 tends to dominate, suppressing deliberate reasoning in favor of quick, heuristic responses. Here, emotions are not merely noise-they act as attention filters, time-compression mechanisms, and motivational signals.[28].

In line with this, according to the Risk-as-Feelings Hypothesis, individuals are more likely to trust emotionally-laden narratives that are familiar, vividly presented, easy to imagine, or connected to personal experience. The more emotional-laden a decision and potential outcomes are, the less people rely on probabilities and rational, deliberative thinking (System 2) while evaluation possibilities, letting their feelings and intuition (System 1) take control (Loewenstein et al., 2001; Slovic et al., 2013).

Barrett (2017a, 2017b) deepens this view by proposing that emotions are not automatic reactions but predictions, which guide human behavior. They are interpretations of internal and external cues, constructed by the brain based on prior experience, memory, context, and social learning. In our context, this implies that investor sentiment is not just a response to markets - it actively constructs meaning from signals such as price movements, news, and macroeconomic data. Often this happens at a subconscious level, thus making even the investor think he (or she) thinks deliberative and acts rational. Closely related to Kahneman (2011)'s System 1, emotions guide attention and interpretation, often reinforcing a worldview or imagined future.

In a somewhat similar vein, Tuckett (2011) introduces the concept of conviction narratives - emotionally charged mental simulations of future scenarios. These narratives function as tools for planning and action in situations where probabilistic reasoning is insufficient due to fundamental uncertainty. From this perspective, investors often rely on narrative templates - scripts of how events may unfold - that help them "feel out" plausible futures.

[27] If deviations from these definitions occur in the studies discussed below, they will be explicitly noted.

[28] As Kahneman himself highlights several times during his book, System 1 and System 2 are theoretical constructs rather than distinct physical systems in the brain.

These templates are not just representations of beliefs, they anchor emotions that make a particular course of action feel right.

In the context of finance, studies have shown that emotions influence what investors attend to, how they interpret information, and which conclusions they draw. Emotionally salient inputs are perceived as more credible and are more likely to be remembered, thereby reinforcing motivated reasoning, selective exposure, and ultimately, belief polarization. These mechanisms collectively undermine the rational ideal of adaptive expectation formation (Hirshleifer, 2015).

All this lines up with Kahneman (2011, p. 129 ff.), who argued that rather than responding purely to data, people navigating hype cycles or disruptive innovations may act based on how a narrative feels. A narrative gains power not through its accuracy, but because it provides a compelling emotional story that aligns with identity and evokes danger, opportunity and urgency. Moreover, narratives that appear more plausible may prevail over those that are statistically more probable, as plausibility is often shaped by individual experiences and reinforced by salience in public discourses.

In addition, it has been shown that specific emotions systematically influence cognition and decision-making. While fear increases caution and risk aversion, anger leads to overconfidence and blame attribution, and excitement heightens perceived opportunities while suppressing recognition of downside risks. Such emotional salience can again enhance memory and subjective confidence - even when predictive accuracy declines (Brosch et al., 2013; Lerner et al., 2015).

Visano (2002) illustrates how emotionally contagious narratives - particularly those linked to identity, status, or group affiliation - generate social and psychological pressures to act. The well-known phenomenon of Fear of Missing Out (FOMO) exemplifies this dynamic: it is not merely a rational anticipation of foregone returns, but a socially leveraged impulse to also invest that suppresses contrarian deliberation in times of stock market bubbles.[29] Narratives common to economic booms often gain traction not because of their analytical robustness, but because they provide emotionally compelling behavioral scripts that appear actionable. Importantly, narratives are not incidental byproducts of human behavior, rather, they are fundamental to it, as they generate meaning and guide action (Taffler et al., 2024).

Closely related to group-related FOMO is Kindleberger and Aliber (2005)'s emphasis on social comparison and Veblen (2017)'s concept of conspicuous consumption[30]: during stock market booms,[31] retail investors observe friends, neighbors and family members accumulating wealth through speculation. In an effort to keep up, they also enter the market - even with inferior information.[32]

[29] FOMO describes the tendency to engage in a behavior primarily because others do so, reflecting humans' social nature and aversion to being left behind. In the context of our theme, this may manifest as investors purchasing NVIDIA stock simply because others are doing so and out of fear of missing potential profits *in comparison to others*.

[30] The idea that people purchase and consume goods primarily to signal wealth and status

[31] While the authors do not provide a formal definition of a boom, they implicitly describe it as a period of stock prices rising well beyond normal fluctuations, such as the surge in internet stocks during the 1990s.

[32] A similar mechanism can increase private indebtedness during downturns, as households seek to maintain consumption levels to match those of their peers.

Beyond this, even experts tend to estimate probabilities and potential rewards in a biased way. Contrary to real-world behavior, experimental studies have shown that even renowned experts tend to evaluate risks within their own domain of expertise inversely to potential returns - perceiving high-reward scenarios as less risky and low-reward scenarios as more risky (Finucane et al., 2000; Slovic et al., 2013).[33]

As emphasized by Shiller (2020), narratives rarely act in isolation but rather in constellations. Although he does not conduct formal statistical modeling, his discussion of the Great Depression (ibid., pp. 136 ff.) illustrates this interplay vividly. He introduces the *Frugality Narrative*: during the 1930s, people adapted to hardship by reshaping their stories and consumption habits. The conspicuous consumption of the 1920s (e.g., buying cars as status symbols) gave way to demonstrative modesty. A form of "poverty chic" emerged, reflected in the widespread adoption of Levi Strauss' jeans and the deliberate damaging of new garments to convey frugality. Thus frugality became a moral virtue, and poverty was re-framed as largely beyond individual control. These narratives fostered precautionary saving and reduced spending. While such behavior may appear rational individually, Shiller argues that it was driven less by calculation than by stories of peers and social contagion. Many also avoided visible consumption out of empathy or fear of social disapproval.[34].In sum, these interlinked narratives - fear of poverty, modesty as virtue, and moral restraint - according to Shiller reinforced the downturn by depressing aggregate demand and thereby prolonging the Great Depression.

More broadly, Shiller (2020) emphasizes that (i) narratives need not be frequently repeated to remain influential, (ii) individuals are often unaware of how narratives subconsciously shape their decisions, and (iii) narratives can mutate and acquire new meanings in changing contexts. Regarding (i), he notes that even during the Great Depression, people did not constantly speak about their fear of losing their jobs, yet this fear continued to influence behavior. An illustrative example of (iii) can likewise be drawn from that period: jeans - originally a symbol of frugality - later came to represent "youthful rebellion" (ibid. p. 148) in the 1940s and women's emancipation in the 1950s.
In sum, a wide range of psychological and social mechanisms may systematically influence investors' behavior. The key question is whether these influences are large and persistent enough that rational arbitrage cannot fully offset them. The frameworks discussed - especially the Risk-as-Feelings hypothesis and Kahneman's dual-process theory - suggest that information is not weighted by objective content alone but by its emotional salience and temporal immediacy. Emotionally compelling narratives, such as those promising extraordinary wealth through investing in AI, can trigger intuitive, affect-driven reactions (System 1) that precede deliberate reasoning (System 2).

Although the existence of System 2 implies that reflection and arbitrage are possible, in practice deliberation often follows rather than corrects initial intuitive judgments. Combined with motivated reasoning, selective exposure, affect-based risk perception, and group dynamics such as FOMO, these processes foster the formation of emotionally charged narratives that amplify market movements. In aggregate, this points toward systematic overreaction to sign-consistent news - where positive narratives fuel excessive optimism and negative ones deepen pessimism - especially when information is conveyed through emotionally resonant stories that reinforce existing narrative constellations.

[33]While high-risk, low-reward and low-risk, high-reward options do exist, the former are typically disregarded and the latter are rarely contested.

[34]Neither motive is inherently irrational: economics does not prescribe how preferences are formed.

Combining this discussion with the formal models from the previous section - namely Hong and Stein (1999)'s underreacting news-watchers and trend-chasing momentum traders, and Daniel et al. (1998)'s overconfident investors - all behavioral frameworks converge, at least partly, on the expectation of short-term same-sign reactions to news, as stated in hypothesis H1, followed by subsequent reversals as postulated in hypothesis H2.[35]

Ultimately, whether such mechanisms manifest in measurable sentiment dynamics and asset prices is an empirical question. The following section therefore reviews key evidence on how emotions and narratives are reflected in market behavior and asset valuation.

3 Theoretical Background: Sentiment inference via FinBERT

Understanding FinBERT-and its underlying base model, BERT - is best achieved by examining three key building blocks: Deep Embeddings, Shallow Neural Networks, and the Attention Mechanism. These components are discussed in the following three subsections, while Section 3.4 focuses on the models themselves.

3.1 Deep Embeddings

Earlier models - such as the dictionary-based approaches discussed above - represented documents using a so-called Bag-of-Words model (BoW) model. Here, a document was encoded as simple word counts, i.e., matrices capturing how often each word from a fixed vocabulary appeared in a given text.[36] Individual words were represented as one-hot vectors: vectors of vocabulary length in which a single position corresponding to the word was set to one and all others to zero.[37] A key limitation of such representations is that they encode no semantic or syntactic relationships between words.

Beginning in the early 2010s, so-called neural embeddings were introduced (A. H. Huang et al., 2023). Unlike BoW, these methods embed each word into a dense, relative to BoW low-dimensional vector space that captures latent semantic and syntactic properties. For an embedding vector X, consisting of elements x_i, the positions no longer correspond to individual words but to latent dimensions inferred during training. Although these dimensions are not directly human-interpretable, they allow algorithms to capture rich

[35] While this may appear straightforward, it poses a theoretical challenge: Hong and Stein (1999) predict that prices rise the next day because agents adjust too slowly, whereas Daniel et al. (1998) attribute the same outcome to investors reacting too strongly and later seeking confirmatory information. Both mechanisms imply short-term momentum but rest on distinct psychological foundations. This highlights that behavioral modeling involves choosing among multiple plausible cognitive biases rather than rejecting one model in favor of another - an inherent complexity when bridging psychology and market dynamics.

[36] In NLP, the *vocabulary* refers to the set of unique tokens representing words, sub-words, or characters that a model can recognize. A given vocabulary is therefore used to translate human-readable input into machine-readable vectors and to convert the model's output back into natural language. Importantly, a model can only 'understand' human language which is in its vocabulary (Mielke et al., 2021).

[37] For instance, if the vocabulary comprised 5,000 words and the 300th word was "cat", then "cat" was represented by a vector of length 5,000 with all entries equal to 0 except the 300th, which was 1.

linguistic structure. Crucially, semantic similarity is reflected in proximity in the vector-space: words with related meanings are located near each other, while antonyms often occupy opposing directions. Moreover, simple vector arithmetic produces meaningful analogies, e.g., king − man + woman ≈ queen (Allen & Hospedales, 2019).

Despite their success, static embeddings face major limitations. They cannot capture polysemy, i.e., the same vector represents "bank" in both "I sat on a bank" and "I worked for a bank." Furthermore, they cannot handle out-of-vocabulary words absent from the training corpus.

A major breakthrough came with the introduction of Transformers (Vaswani, 2017) and Large Language Model (LLM)s such as BERT. These models generate *contextual embeddings*, in which word vectors are built dynamically, conditioned on surrounding tokens. In addition, modern tokenization operates at the sub-word level (e.g., WordPiece), enabling the representation of morphological variants and previously unseen words (Mielke et al., 2021). For example, the tokens *man* and *'s* can be combined to represent *man's*. This not only improves handling of plurals and misspellings but also allows models to robustly process unknown terms. Most importantly, contextual embeddings distinguish subtle semantic differences: they can differentiate not only between "I sit on a bank" and "I worked for a bank," but also between "I sit on a bank" and "I sit on a bank with my girlfriend."

To appreciate how such context-sensitive representations are possible, it is necessary to understand Artificial Neural Network (ANN)s and the Attention Mechanism, which will be discussed in the following sections.

3.2 Artificial Neural Networks

ANNs, often referred to as feed-forward networks, are foundational for many modern Machine Learning architectures (Goodfellow et al., 2016, S. 159 ff.), including Transformers (Vaswani, 2017).

The basic principle is straightforward: a set of (potentially high-dimensional) input features is subjected to successive non-linear transformations to produce an output.[38] In supervised learning, during training,[39] this output is compared with the true values of the target variable, and the discrepancy is measured through a loss function. The model parameters are then iteratively updated to minimize this loss via backpropagation.[40] This cycle of prediction and parameter adjustment is repeated until the loss converges to an acceptable level or ceases to improve.

Alongside these trainable parameters, ANN also contain *hyperparameters*, which are chosen by the modeler before training begins (e.g., the number of layers or the form of the loss function, see below). The quality of training outcomes depends heavily on these choices.

[38] In machine learning terminology, input features correspond to what econometricians call independent variables or predictors, while the target variable corresponds to the dependent variable.

[39] There also exist learning variants without known true values of the target variable. In such cases, unsupervised learning - as discussed for Latent Dirichlet Allocation (LDA) - is applied. A detailed treatment of this is beyond the scope of this thesis. For further discussion see Goodfellow et al. (2016).

[40] In brief, backpropagation involves two steps: (i) computing the gradient of the loss function with respect to each model parameter, and (ii) applying an optimization algorithm such as Adaptive Moment Estimation (Adam) to adjust the parameters in the opposite direction of the gradient. For a detailed discussion, see Hecht-nielsen (1992).

Optimizing them - referred to as *hyperparameter tuning* - involves repeatedly training the model under different hyperparameter settings, evaluating its performance, and selecting the configuration that yields the best results (Goodfellow et al., 2016, S. 373 ff.).

Once trained, the model can then be applied to make inferences on previously unseen data.

The main advantage of ANNs over classical econometric models - and even over many traditional Machine Learning approaches - is their ability to capture highly complex, non-linear inter-dependencies between a large set of variables. This property is theoretically underpinned by the Universal Approximation Theorem, which states that even a simple neural networks with sufficient computational capacity (i.e., a sufficiently large number of neurons, see below) can approximate *any* continuous function (Hornik, 1991).[41]

Diving into architectural detail, an ANN is composed of three types of layers: an input layer that receives the input features, one or more hidden layers where most of the computation takes place, and an output layer that produces the final prediction.

Each layer is made up of simple computational units called *Neurons*. A Neuron receives input values, multiplies them with weights, adds a bias term that shifts the signal, and then applies a non-linear activation function, such as the Sigmoid.[42] Mathematically, the computation performed by hidden layer l with n_l neurons can be written as

$$\mathbf{Z}^l = \mathbf{W}^l \mathbf{A}^{l-1} + \mathbf{b}^l, \qquad \mathbf{A}^l = g(\mathbf{Z}^l) \tag{13}$$

$\mathbf{A}^{l-1} \in \mathbb{R}^{n_{l-1} \times N}$ denotes the activations from the previous layer (with $\mathbf{A}^0 = \mathbf{X}^0 \in \mathbb{R}^{n_0 \times N}$ representing the input data),[43] N the number of observations, $\mathbf{W}^l \in \mathbb{R}^{n_l \times n_{l-1}}$ is the weight matrix, $\mathbf{b}^l \in \mathbb{R}^{n_l \times 1}$ is the bias vector (broadcasted to $\mathbb{R}^{n_l \times N}$ during computation),[44] and $g(\cdot)$ is a non-linear activation function applied element-wise. The resulting activations are $\mathbf{A}^l \in \mathbb{R}^{n_l \times N}$.[45]

The specific activation function, the number of Neurons in a given layer as well as the number of total hidden layers are hyperparameter.

The output layer applies a final activation function, which can be written as

$$\hat{\mathbf{Y}} = f(\mathbf{A}^d) \tag{14}$$

[41] In practice, a neural network cannot perfectly learn any arbitrary function. There are three main reasons for this: i) there is no guarantee that the optimization algorithm will find the globally optimal set of parameters, ii) the model may suffer from over-fitting, i.e. from just memorizing the training-data (including noise), and iii) the theoretically required computational capacity may be prohibitively large (Goodfellow et al., 2016, S. 192 ff.).

[42] One might thus even consider a Logistic Regressions as the most basic form of a Neural Networks.

[43] The number of neurons in the input layer equals the number of input features, i.e., each input feature is represented by one neuron.

[44] Broadcasting refers to the computational process of taking an $n \times 1$ vector and automatically expanding it across N training examples, so that each of the n bias terms is applied simultaneously to all N observations. This avoids explicit duplication and enables efficient vectorized computation. It is handled automatically by modern libraries such as NumPy for Python (NumPy, n.d.).

[45] Common activation functions include the Sigmoid, Hyperbolic Tangent, or the Rectified Linear Unit (Ertel, 2021; Terven et al., 2023).

Here, the activation function f is a hyperparameter selected according to the prediction task. For instance, a softmax function is commonly used for multi-class classification (Kipp, 2024a, 2024b). [46]

The power of ANNs does not stem from a single neuron - whose computations are mathematically trivial - but from stacking many neurons across multiple layers. This layered structure enables the network to learn increasingly complex representations of the data. Non-linear activation functions are crucial in this process: without them, even deep networks would collapse into a single linear transformation and would be no more powerful than ordinary linear regression (Kipp, 2024b).[47]

To illustrate the computational process, Figure 2 depicts a simple ANN with three input features, one hidden layer of three neurons, and three output neurons. Each arrow represents a weight connecting two neurons, while each neuron applies its own bias. For clarity, not all weights and no biases are drawn explicitly. The dashed arrows highlight selected weights, but in a real model every neuron is connected to every neuron in the subsequent layer, making the network a *fully connected* architecture.

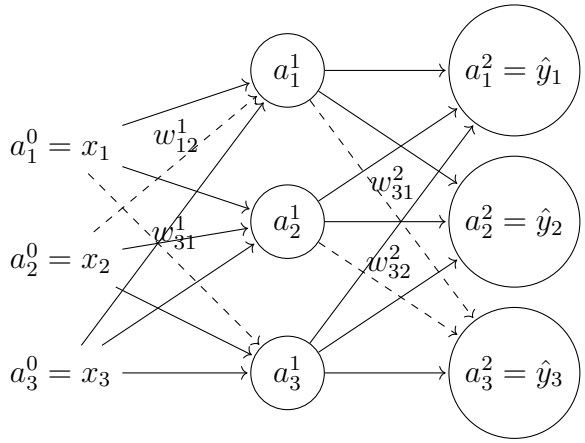

Figure 2. Structure of a simple ANN with one hidden layer (bias terms omitted). The variable a_i^l denotes the activation of the i-th neuron in layer l. Dashed lines highlight selected weights. *Source:* Own illustration based on Dony and Haykin (1995) and Stutz (2020).

Note that in modern applications networks are typically much deeper, and a network with only one or a few layers is often described as *shallow*. In practice, one or several layers of an ANN are frequently combined with other architectures, such that the ANN becomes part of a larger model. In these contexts, they are commonly referred to as *feedforward layers* or *linear layers*, though the underlying computations remain identical to those described above (Sarker, 2021).

With this brief foundation in mind, we will move to the other key-component of Transformers: the Attention mechanism.

[46] Formally, the output layer can also include weights and biases, making its structure analogous to that presented in Equation 13.

[47] The *depth* of a network refers to the number of trainable layers, where the input layer is usually not counted (Kipp, 2024b).

3.3 The Attention Mechanism

The attention mechanism was introduced as a method for dynamically weighting information across a sequence of text-encoding tokens, enabling the model to emphasize the most relevant aspects of a given input (Bahdanau et al., 2014; Vaswani, 2017). The key innovation is that, rather than interpreting an embedding-vector x_i in isolation as a fixed (neural) embedding, the model can contextualize it by weighting in neighboring embeddings along the whole sequence contained in X.[48]

To compute -product attention,[49] the input is first projected into three distinct subspaces, producing the matrices of *queries* (Q), *keys* (K), and *values* (V). This allows the model to represent what each token is "asking for" (Q), what it "answers" (K), and the information it contributes (V).[50]

Formally, let $X \in \mathbb{R}^{N \times H}$ denote the sequence of N token embeddings of size H. Applying learned projection matrices yields:

$$Q = XW^Q, \qquad K = XW^K, \qquad V = XW^V \tag{15}$$

where $W^Q, W^K \in \mathbb{R}^{H \times H_k}$ and $W^V \in \mathbb{R}^{H \times H_v}$ are trainable weight matrices, and consequently $Q, K \in \mathbb{R}^{N \times H_k}$ and $V \in \mathbb{R}^{N \times H_v}$.

Next, similarity scores are computed via the dot product $QK^\top$,[51] rescaled by $\sqrt{H_k}$ (the dimensionality of the keys) to control magnitudes, and passed through a softmax to obtain the *attention weights*, which determine the contribution of each value vector in V:

$$\text{Attention}(Q, K, V) \;=\; \text{softmax}\left(\frac{QK^\top}{\sqrt{H_k}}\right) V \tag{16}$$

Multi-head attention repeats this mechanism in parallel across A heads (typically with $H_k = H/A$ per head), and then concatenates and projects the results back to the model dimension:

$$\begin{aligned} \text{MHA}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_A) W^O, \\ \text{head}_i &= \text{Attention}(QW_i^Q,\, KW_i^K,\, VW_i^V). \end{aligned} \tag{17}$$

Empirically, different heads often specialize in distinct linguistic aspects (e.g., syntactic dependencies, semantic relations, positional patterns). Interestingly, these roles are not programmed by the modeler, but instead emerge naturally during training (Clark et al., 2019).

[48] The definition of "whole sequence" differs between architectures. In auto-regressive models (Radford et al., 2018, e.g.), attention is restricted to preceding elements of the sequence (i.e., for some $x_i \in X$, only tokens x_{i-j} with $j \geq 0$ are considered). By contrast, bidirectional models permit attention over the entire sequence, including x_{i+j} with $j \geq 0$ (Devlin et al., 2019; Vaswani, 2017).

[49] Here the focus is on scaled dot-product attention, as it is the core mechanism used in Transformer architectures (Vaswani, 2017).

[50] Intuitively, For a given token A, its query vector Q_A asks all tokens in the sequence how relevant they are for its output. Each of those tokens replies with its key vector K. If token A is particularly relevant for token B, the dot product between Q_B (B's query) and K_A (A's key) will be large, giving A's value V_A a high weight in B's output.

[51] These are raw dot products. Some models normalize Q and K before applying attention, which turns the measure into cosine similarity.

In short, self-attention preserves ANNs ability to learn complex functions, while drastically improving the ability to account for context (Clark et al., 2019; Vaswani, 2017).

3.4 BERT and FinBERT

Building on the attention mechanism (Bahdanau et al., 2014) and especially the Transformer architecture introduced by Vaswani (2017), Devlin et al. (2019) developed Bidirectional Encoder Representations from Transformers (BERT), which is one of the most widely applied models in modern NLP (Acheampong et al., 2021). At its architectural core, BERT is composed of a stack of identical encoder blocks,[52] each containing two main components: (i) a multi-head self-attention layer, and (ii) a position-wise feed-forward network.

While (i) was discussed in the previous section, (ii) can be understood as a slightly adapted version of the ANN introduced in Section **??**.[53] The attention sublayer enables each token receive information from all other tokens in the sequence, while the feed-forward sublayer enriches the representation of each token itself. In this way, the two mechanisms complement each another: attention captures relationships *between* tokens, while feed-forward layers refine the information *within* tokens.

Both sublayers are wrapped with residual (or *skip*) connections[54] and layer normalization,[55] which together improve stability during training. Since the attention mechanism does not capture word order directly,[56] positional encodings are added to the input embeddings to retain information about token positions.[57]

Mathematically, a single encoder block in BERT can be written as

$$\begin{aligned} \mathbf{Z} &= \text{LayerNorm}\big(\mathbf{X} + \text{MHA}(Q, K, V)\big), && \mathbf{Z} \in \mathbb{R}^{n \times H}, \\ \mathbf{H}' &= \text{LayerNorm}\big(\mathbf{Z} + \phi(\mathbf{Z}W_1 + b_1)W_2 + b_2\big), && W_1 \in \mathbb{R}^{H \times H_{ff}},\ W_2 \in \mathbb{R}^{H_{ff} \times H}, \\ & && b_1 \in \mathbb{R}^{H_{ff}},\ b_2 \in \mathbb{R}^{H},\ \ \mathbf{H}' \in \mathbb{R}^{n \times H}. \end{aligned} \tag{18}$$

Here, $\mathbf{X}$ is the input to the encoder block, $\text{MHA}(Q, K, V)$ denotes the multi-head attention output, and ϕ the Gaussian Error Linear Unit (GELU) activation (Devlin et al.,

[52] Encoder architectures transform an input X into deeper embeddings, whereas decoders map embeddings back into human-interpretable outputs. Since BERT is an encoder-only model, a task-specific output layer must be added (see below).

[53] Concretely, it consists of two fully connected layers with a non-linear activation in between.

[54] Skip connections add the input of a sublayer directly to its output, which stabilizes training and prevents information loss across deep networks (He et al., 2016).

[55] Normalization here is analogous to computing z-scores in econometrics, i.e., centering each input around zero and scaling by its standard deviation. In Transformers, this is performed per token across its feature dimensions, ensuring stable distributions.

[56] Classical models for sequence learning, like Long Short-Term Memory (LSTM), take the inputs sequentially at each time step. While this directly incorporates positional information, it also slows performance as it hinders parallel computation (Kipp, 2024c).

[57] In BERT, positional information is added through trainable positional embeddings, which are summed with the token embeddings at the input layer.

2019; Vaswani et al., 2017).[58] The matrices W_1 and W_2 parameterize the feed-forward layer with a dimensionality of $H_{ff} = 3072$,with b_1, b_2 as bias terms. LayerNorm$(\cdot)$ refers to layer normalization, and residual connections are made explicit by the right-hand side of each equation. In the BERT-Base model, $H = 768$, $L = 12$, and $A = 12$, resulting in approximately 110 million trainable parameters.[59]

As illustrated in Figure 3 below, each input sequence is first tokenized using the WordPiece tokenizer (Wu et al., 2016). A [CLS] token is added at the beginning for sequence-level classification, and a [SEP] token marks the end of the input. If two sequences are provided, an additional [SEP] separates them. Importantly, model input per pass cannot vary from the model's fixed input length (512 tokens), ensured via truncation or padding.[60]

BERT then constructs the final input embeddings by summing three components: the token embeddings, trainable positional embeddings $E_{\text{position}} =$, and segment embeddings E_{segment} (the latter only distinguishing between sentence A and B in paired inputs).[61][62]

The resulting embeddings are processed by L stacked encoder layers with A attention heads. The final hidden state of [CLS] serves as a sequence-level representation, which is passed to a task-specific output layer (e.g., a softmax classifier).[63]

Figure 3 provides a simplified schematic of the entire model, showing how embeddings are transformed into contextual representations through stacked encoder blocks.

[58] The GELU, introduced by Hendrycks and Gimpel (2016), is defined as $\text{GELU}(x) = x\,\Phi(x)$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Unlike Rectified Linear Unit (ReLU), which applies a hard threshold at zero, GELU weights inputs by their probability of being positive. This yields smoother gradients, improves training stability, and has proven effective in deep Transformer architectures. Since computing $\Phi(x)$ is costly, BERT employs a tanh-based approximation: $\text{GELU}(x) \approx 0.5\,x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\,(x + 0.044715\,x^3)\right)\right)$.

[59] In another model-specification, called BERT-Large, Devlin et al. (2019) increases H to $H = 1024$, L to $L = 24$, the number of attention heads to $A = 16$ and the conditionality of the feed-forward sub-layers to $H_{ff} = 4096$. In total, this yields roughly 340 million trainable parameters.

[60] Truncation refers to artificially shortening the input by dropping tokens, while padding refers to adding zeros to the input until the limit is reached.

[61] Unlike W^Q, W^K, W^V, the embedding tables $E_{\text{token}}, E_{\text{position}}, E_{\text{segment}}$ are themselves trainable weight matrices. They are updated during pre-training through back-propagation and do not require an additional projection matrix W^E (Devlin et al., 2019; Research, 2020).

[62] All embeddings are always part of BERT's input. For single-sequence tasks, however, all tokens receive the same segment ID, functioning like a single bias-term, shifting the whole distribution without adding additional information. In sentence-pair tasks, however, tokens are distinguished by segment IDs, which shifts the distributions in different sentences differently.

[63] Clark et al. (2019) demonstrate that attention heads treat the [SEP] token as a placeholder when they do not contribute meaningful information in a given context.

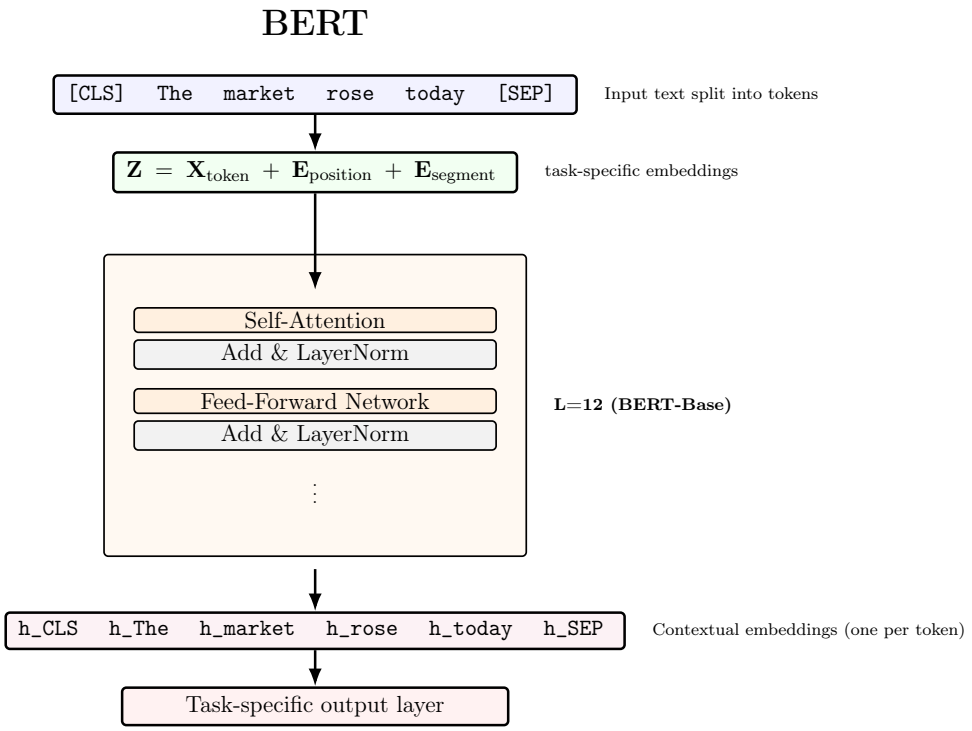

Figure 3. Simplified schematic of the BERT model. Equation 18 shows the computations inside each encoder block, which are stacked repeatedly (BERT-Base: 12 blocks) to generate contextual token representations. *Source:* Own illustration, based on Devlin et al. (2019) and utilizing generative AI (https://chatgpt.com/g/g-p-67b49e1e74848191b4b1cda5221e7191/c/68dceaca-67dc-832d-a0bb-83d77ba2e69f).

Besides its architecture, a key innovation of BERT lies in its pre-training strategy, which enables the model to learn contextual embeddings from large unlabeled text corpora.

This is achieved through two pre-training objectives:

1. *Masked Language Modeling (MLM):* 15% of tokens are randomly selected for prediction. Of these, 80% are replaced ("masked§) with the special [MASK] token, 10% with a random token, and 10% remain unchanged.[64] The model is trained to recover the original token, forcing it utilize information from both left and right context.

2. *Next Sentence Prediction (NSP):* The model receives pairs of sentences and must predict whether the second follows the first in the original text (50% of the time this is true). This task forces the model to learn dependencies across sentences.

Those tasks were performed on a large, general-domain corpora (BookCorpus and English Wikipedia, totaling ≈ 3.3 billion words), learning complex semantic and syntactic dependencies. These representations can be adapted through *fine-tuning*: Here the pre-trained model is further optimized for specific tasks, such as filtering our relevant sections of sustainability reports (Luccioni et al., 2020) or classify patents (J.-S. Lee & Hsiang, 2020).[65]

[64] This strategy prevents the model from overfitting to the [MASK] token and encourages more robust context-based representations.

[65] In machine learning, this is called *transfer learning*: a model pre-trained on massive corpora with substantial computational resources acquires rich embeddings, which can later be adapted with comparatively modest amounts of data and hardware to perform well on downstream tasks (Zhuang et al., 2020).

For both pre-training and fine-tuning, BERT minimizes a task-specific loss (e.g., cross-entropy for classification[66]) using the AdamW optimizer with warmup and subsequent linear learning-rate decay (Loshchilov & Hutter, 2017).[67],[68] η is the step size, and λ the weight decay (Loshchilov & Hutter, 2017). Conceptually, the loss quantifies the deviation between prediction and true values, while the optimizer alliteratively adjusts the parameters by changing them into the opposite direction of the gradient.[69]

A fine-tuned BERT-model is FinBERT, introduced by Araci (2019), which is optimized in modeling the specific syntactic and semantic nuances of financial language.

While preserving the same architecture, FinBERT was further trained on financial texts to better model the nuances of this domain. For domain adaptation, the authors compiled a corpus of 46,143 financial news articles (amounting to more than 29 million words) and continued masked language modeling to adjust BERT's embeddings towards finance-specific vocabulary and usage patterns.

For supervised fine-tuning, the authors used the Financial Phrase Bank developed by Malo et al. (2014). The dataset comprises 4,845 sentences from financial news articles, each annotated for sentiment toward specific stocks. Each sentence was labeled according to whether the news was expected to affect the mentioned stock positively or negatively. The fine-tuning objective was to predict the correct sentiment label via a softmax classifier that took the [CLS] token representation from the final layer as input and produced the sentiment label as output.

Empirical evidence shows that FinBERT outperforms both general-purpose BERT models and traditional machine learning baselines in financial sentiment classification. Moreover, it has been shown to produce robust results when fine-tuned on other finance-related tasks, even with very limited training data (Acheampong et al., 2021; Mosbach et al., 2020; Research, 2020).

Consequently, this thesis employs FinBERT as the primary model for sentiment inference. Before discussing methodologies employed more deeply in section 5, the next paragraph will discuss relevant empirical studies regarding the quantification of emotions, investor sentiment and narratives.

4 Literature review

The focus of the literature review will be on what Gentzkow et al. (2019) termed "Text as Data": the application of NLP to enable the utilization of text in quantitative models.[70]

[66] Cross-entropy for a prediction $\hat{y}$ against the true value y is defined as $L = -\sum_i y_i \log(\hat{y}_i)$.

[67] AdamW updates parameters θ according to $\theta_{t+1} = \theta_t - \eta \cdot \frac{m_t}{\sqrt{v_t}+\epsilon} - \eta \cdot \lambda\theta_t$, where $m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$ and $v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2$ are the exponentially moving averages of the first and second moment estimates of the gradient $g_t = \nabla_\theta \mathcal{L}(\theta_t)$.

[68] Warmup refers to a method in which the optimizers step size is gradually increased during the early training phase, a procedure shown to stabilize optimization. In AdamW, this is combined with later learning-rate decay, i.e. with lowering the step-size in later steps.

[69] One may visualize this as a (loss-) landscape with hills and valleys. Each position here corresponds to a specific setting of the parameters θ. The gradient tells us the steepest way uphill. By stepping in the opposite direction, the parameters will lead the model downhill, finding lower positions at each step.

[70] NLP can be defined as a sub-field of Computer Science that "employs computational techniques for the purpose of learning, understanding, and producing human language content" (Hirschberg & Manning, 2015).

The focus is motivated by the fact that such methodologies outperform traditional measures of investor sentiment and emotions, which relied on noisy, market-based indicators such as trading volume, price volatility, closed-end fund discounts (e.g. M. Baker & Wurgler, 2006), or survey-based indicators (Zhou, 2018).[71] In addition to higher precision, a key strength of the "Text as Data" approach lies in its ability to confront the challenge, emphasized by Black (1986), that many economically relevant variables are intrinsically difficult to quantify (Gentzkow et al., 2019).

An early empirical demonstration of how emotional tone in financial news affects market prices is provided by Tetlock (2007), who showed that pessimistic language in the WSJ's *Abreast of the Market* column predicts lower subsequent stock returns.

Using the so-called General Inquirer (GI), the author transformed all columns for the period 1984-1999 into distinct matrices capturing the presence of 77 emotional categories.[72] To further reduce dimensionality, Principal Component Analysis (PCA) was applied, with the first principal component loading predominantly on pessimistic categories such as Negative, Weak, Fail, and Fall.[73]

Building on the Noise Trader framework and related concepts discussed in Section **??**, the author distinguished three possible channels through which pessimistic media content might affect markets:

- **Trigger channel**: pessimistic media trigger investor sentiment, leading to short-term price drops followed by mean reversion,
- **Reflect channel**: media merely reflect prevailing sentiment, e.g. high pessimism after poor returns,
- **Information channel**: media convey fundamental information, in which case no reversal would be expected.

The first two mechanisms correspond to what Tetlock (2007) terms *sentiment theories*, whereas the latter represents an *information theory*. These can be mapped onto equation (12): when news contain fundamental information (information channel), they affect r_t or D_t, while sentiment-related content influences A_t. Notably, the *trigger channel* closely mirrors the mechanisms discussed in the previous section - namely, the potential for emotionally charged news to induce overreactions driven by System 1 processes.

[71] While surveys might catch narratives and emotions very precisely, they are far too infrequent for granular analysis (Bollen et al., 2011; Shiller, 2021; Zhou, 2018).

[72] According to the author, the General Inquirer is a "well-known quantitative content analysis program" which compresses text into 77 categories of the Harvard-IV Dictionary. By counting how often words related to each category appear in an article, the emotional content is represented as a simple matrix (ibid., p. 114). The Harvard-IV Dictionary in turn is one of the oldest and most widely used sentiment and content analysis lexicons (Price et al., 2012).

[73] In brief, PCA begins by computing the variance-covariance matrix of the variables and deriving its eigenvalues and eigenvectors. Intuitively, eigenvectors define directions in which the data are spread. The eigenvector associated with the largest eigenvalue points in the direction of the greatest spread, and the first principal component is the linear combination of the original variables along this eigenvector, thereby capturing the maximum variance in the data. The second (third) eigenvector corresponds to the second (third) principal component and captures the second (third) largest share of the variance (Greenacre et al., 2022).

To test these channels, Tetlock (2007) estimates Vector Auto-regressions of daily returns and trading volume on lagged values of both variables, augmented by the pessimism measure and some controls (e.g., volatility, weekday and calendar dummies).[74][75]

The results support the trigger channel: stock prices fall significantly in the first one to two days following pessimistic media coverage, but revert to the mean within the next two to five days. Furthermore, trading volume increases conditional on extreme pessimism, consistent with heightened information uncertainty and sentiment-driven trading.

Engelberg and Parsons (2011) also explicitly focused on the way in which media and stock prices may influence each other. More precisely, using brokerage account data from 19 U.S. cities with a total of 15,951 households during the period 1991–1996, the authors studied whether coverage of local newspapers of S&P 500 earnings announcements *causally* affects retail investor trading.

They regressed local trading volume on an indicator variable for whether the local newspaper reported on a firm's earnings announcement within three days of its release, while simultaneously controlling for a large set of potential confounding factors (e.g. geographic proximity to the firm's headquarters, national media coverage, and investor portfolio composition).

The results show that local coverage increased local trading volume by 8–50%, with robust effects for both buying and selling. To establish causality, the authors exploited exogenous variation in media exposure, such as severe local weather events that disrupted newspaper delivery.[76] Consistent with a causal effect from news on stock markets, trading spikes only when coverage actually reached investors, and the effect vanishes when delivery was disrupted.

In another example, Bollen et al. (2011) derived two dictionary-based sentiment measures for Twitter data. First, using OpinionFinder (see Wilson et al., 2005), they constructed a binary index by classifying tweets as positive or negative and aggregating these counts for each day. Second, they built a six-dimensional mood series[77] based on the GPOMS lexicon, which extended the original POMS-bi inventory by mapping its 65 mood terms to 964 semantically related words using Google n-grams. Tweets were then scored according to these categories and averaged to obtain daily values.

Analyzing nearly 10 million tweets from February to December 2008, the authors are able to show that the mood indices respond sensibly to major events as expected, including the U.S. presidential election and Thanksgiving.

For predictability, the authors estimated Granger Causality tests of first differences in the Dow Jones Industrial Average (DJIA) (i.e., short-term returns) on the lagged sentiment measures. For the six-dimensional mood-measure, only calm showed significant predictive power in a linear setting, while the binary indicator failed to show any predictive power.

To account for non-linear dynamics, they additionally employed a Self-Organizing Fuzzy Neural Network. While incorporating calm improved out-of-sample forecast accuracy, adding both calm and happy further reduced prediction error. This suggests that while

[74] The author interprets trading volume as a proxy for information uncertainty.
[75] In these specifications, this is essential a Granger Causality test.
[76] Note that during the sample period, internet access was not yet widespread.
[77] Calm, Alert, Sure, Vital, Kind, and Happy.

calm has both linear and non-linear predictive value, the role of Happiness emerges only in a non-linear setting.

Another example of the application of dictionary-based methods in direct relation to asset prices is provided by Agarwal et al. (2025), who employed two dictionaries covering both Western and Chinese media sources to extract eight distinct emotions, which they used as independent variables in a series of Ordinary Least Squares (OLS) regressions on daily stock prices during two Chinese stock market bubbles (2005-2008 and 2014-2016).

Interestingly, they observed strong clustering in the spikes of negative emotions like panic, revulsion, denial, blame, but not in positive emotions.

Regard stock prices, their results indicate that strong positive emotions (e.g., excitement) are positively correlated with contemporaneous returns, whereas negative emotions (e.g., blame and anxiety) are negatively correlated with contemporaneous returns and positively with volatility. Furthermore, the absolute predictive power of what they refer to as 'strong' emotions (e.g. excitement, anxiety) is significantly above that of 'weaker' emotions (.eg. happy, worry).

In a similar vein, Taffler et al. (2024) constructed their own dictionaries (with 834 emotion-related terms) to capture "exciting" and "anxious" emotions during the Dot-Com Bubble, the Global Financial Crisis, and the COVID-19 pandemic.

Empirically, the authors applied OLS regressions, to analyze contemporaneous monthly stock returns alongside several proxies for uncertainty and captured emotions.[78] Specifically, they regressed monthly returns on emotion measures and risk proxies, each considered separately, and additionally estimate regressions with risk proxies as dependent variables, using returns and emotions as explanatory factors.

The explanatory power of emotions shows conditionality on a given period: during the Dot-Com bubble, measures of excitement and mania show strong contemporaneous correlations with returns, whereas during the Global Financial Crisis these associations weaken and anxious emotions instead load significantly on proxies for uncertainty. A relatively similar pattern holds in the COVID-19 pandemic, where anxious emotions are particularity good predictors for risk proxies.

Interestingly, Taffler et al. (2024) explicitly situate their research within the framework of Shiller (2020), stating that they 'measure the power of economic narratives by employing textual analysis and carefully constructed emotion keyword dictionaries to identify the specific emotions conveyed in these' (ibid., p. 2). Importantly, the authors thus argue that emotions serve as proxies for underlying narratives, rather than narratives deriving their importance from the emotions they contain.

Turning our attention more directly to narratives, two prominent examples that measured economic narratives using the relatively simple method of dictionary-based term-identification are Geopolitical Risk Index (GPR) and the Economic Policy Uncertainty Index (EPU).[79][80]

[78] The emotion variables were constructed as relative word frequencies, calculated by dividing the count of each emotion term by the total word count across all articles within a given month.

[79] Although one might easily interpret these methods as narrative-based, the authors themselves do not explicitly frame their approaches in these terms.

[80] N-grams are contiguous sequences of words: unigrams consist of one word, bigrams of two words, trigrams of three, and so forth. Task-relevant combinations of words can thus be systematically identified.

The central idea of S. Baker et al. (2016)'s EPU is to quantify economic policy uncertainty through newspaper coverage. Specifically, they examined ten leading U.S. newspapers and counted articles that jointly contain (i) a term related to uncertainty, (ii) a term referring to the economy, and (iii) one of six policy-related terms (Congress, deficit, Federal Reserve, legislation, regulation, or White House).[81] For each newspaper and month, the number of matching articles was scaled by the total volume of published articles to control for size differences. The resulting series were standardized, averaged across newspapers, and normalized to form a composite index.

Despite the methodological simplicity compared to modern NLP approaches (Mishev et al., 2020, see), the index has proven highly informative: it co-moves with other measures of economic uncertainty, spikes around major political and economic shocks, and is widely used both in academic research and by professional data providers such as Bloomberg, FRED, Haver, and Reuters (S. Baker et al., 2016, p. 1595). Importantly, the construction of the policy term set was not ad hoc, it was derived from an extensive human audit of 12,000 articles, in which trained research assistants tested and refined candidate terms through systematic validation exercises.

Methodologically related, the GPR index by Caldara and Iacoviello (2022) measures geopolitical tensions using newspaper text. The authors constructed a hand-crafted dictionary of bi-grams and trig-rams and identify relevant phrases through frequency counts.[82] To refine the measure, the raw counts were weighted using Term Frequency–Inverse Document Frequency (TF–IDF),[83] thereby emphasizing expressions that are both distinctive and contextually salient.

The index exhibits sharp increases during major historical events, e.g. after 9/11 and the World Wars, and has been shown to be a reliable predictor of subsequent declines in investment, stock prices, and employment, which the authors explain by the associated uncertainty measured (Caldara & Iacoviello, 2022).

Besides dictionary-based studies, some authors directly measure narratives by applying more advanced NLP-techniques (Mishev et al., 2020, see).

Bybee et al. (2020) analyzed about 800,000 WSJ articles (1984-2017) using LDA to uncover latent narratives in business news.[84] Their topic model yields interpretable themes whose relative prevalence captures how media attention to economic issues evolves over time. Most topics display persistence, while others, such as "recession" or "panic," cluster around major shocks.

The authors showed that these news-based topics not only mirror established economic indicators but also contain predictive content. Embedding topic attention in vector au-

[81] The newspapers are USA Today, the Miami Herald, the Chicago Tribune, the Washington Post, the Los Angeles Times, the Boston Globe, the San Francisco Chronicle, the Dallas Morning News, the New York Times, and the WSJ. The policy terms include common variants such as "the Fed" or "regulatory."

[82] Although the approach may appear straightforward, the authors invested considerable effort in validation, including manual auditing of 44,000 front pages of The New York Times.

[83] TF–IDF is a standard metric in NLP. For a given term t in a document d, it is defined as the frequency of t in d relative to its average frequency across all documents in the corpus. The underlying intuition is that terms frequent in one document but rare across others carry higher informational value.

[84] LDA is an unsupervised dimensionality-reduction technique which represents each document as a set of latent topics, and each topic as a probability distribution over words. It infers these latent topics from the co-occurrence of words, thereby projecting the high-dimensional text into a small subspace of humanly-interpretable topics (Blei et al., 2003). Unsupervised refers to the fact that the model is not trained on predefined labels which represent known topics. Instead, it discovers patterns on its own.

toregressions, they find that recession-related coverage forecasts declines in output and employment even after controlling for standard financial and macro variables.

In a similar vein, Mai and Pukthuanthong (2021) employed a semi-supervised LDA approach to extract ten predefined economic narratives (e.g., Panic, Bubbles, Real Estate), based on Shiller (2020), from over 7 million New York Times articles (since 1871) and 2 million WSJ articles (since 1889).[85]

Running a series of vector auto regressions, the authors demonstrated that narrative-based measures significantly forecast excess stock returns and volatility at the one-month horizon, with predictive accuracy strengthening after 2000 and exceeding that of conventional macroeconomic (e.g. the VIX, inflation) and sentiment (e.g. the EPU) indicators.

For the New York Times, the most-predictive indicator refers to panic, which positively forecasts returns and negatively forecasts volatility. According to the authors and interpreted with ICAPM, panic functions as a state variable capturing time-varying risk aversion: heightened panic raises the required risk premium while simultaneously reducing investors' willingness to bear risk.

In relation to the WSJ, the strongest one is 'Stock Market Bubble', which is inversely correlated with next-months excess-returns, implying mean-reversion, consistent with models discussed above.

Interestingly, Mai and Pukthuanthong (2021, p.4) also controlled for real-world events - such as recessions, bank failures, wars, disasters, and epidemics - via dummy variables and still find that the Panic narrative retains significant predictive power, with stronger loadings than the dummies themselves. This suggests that it is the narrative itself, rather than the occurrence of the events, that drives excess returns.

Another strand of the literature examines the heterogeneous impact of sentiment across different types of stocks. M. Baker and Wurgler (2006) documented that securities which are more difficult to value and harder to arbitrage are particularly vulnerable to sentiment shocks. Similarly, Brown and Cliff (2004) found that small-cap returns display some sensitivity to changes in sentiment, although the magnitude is relatively modest.[86] Consistent with these findings, Li et al. (2014) showed that sentiment extracted from Chinese financial discussion forums has greater predictive power for small-cap stocks than for larger firms.

Finally, a growing body of literature employs state-of-the-art Transformer-based language models - such as BERT, FinBERT, and GPT - to capture emotions and sentiment in financial texts (see section 3 for a discussion of this model class and section **??** for details on BERT and FinBERT). In a comprehensive review, Acheampong et al. (2021) surveyed studies applying various Transformer architectures to estimate investor sentiment. While no universally accepted best-practice approach has yet emerged, variants of BERT currently appear to yield the most consistent and promising results.

[85]Semi-supervised refers to the fact that seed words are given to the model to ensure the convergence of extracted topics toward pre-selected narratives. Or, to put it simply, it is made sure that the model focuses on pre-selected topics, rather than inferring them without any guidance.

[86]Compared to most subsequent work, Brown and Cliff (2004) are cautious in their assessment of sentiment's explanatory power. This may partly reflect their reliance on indirect sentiment proxies such as NYSE turnover and the number of Initial Public Offering (IPO)s, which are weaker measures relative to modern NLP-based approaches (Zhou, 2018).

Niţoi et al. (2023) developed a Central Bank Sentiment Index by fine-tuning BERT on manually labeled sentences from central bank minutes of Czechia, Hungary, Poland, and Romania. The objective was to capture the monetary policy stance conveyed in board discussions, and to evaluate whether sentiment provides information about future policy decisions beyond standard macroeconomic indicators.

The dataset covered 591 policy minutes (50,451 sentences) between 2007 and 2022. From these, 1,998 sentences were randomly sampled and manually annotated as either hawkish, neutral, or dovish, corresponding to restrictive, balanced, or expansionary stances. Using a 90/10 train-validation split, the authors fine-tuned BERT and achieved an F1-Score of 0.88, indicating strong classification performance.[87]

The model assigned each sentence a hawkish (+1), dovish (−1), or neutral (0) label and constructed document-level scores by averaging across sentences. Applied to the full corpus, this procedure generated country-level time series, which were subsequently evaluated in ordered probit models of policy rate decisions while controlling for inflation, output gaps, and exchange rates. Across all four countries, the BERT-CBSI proved to be a robust and superior predictor of future policy rate changes compared to traditional dictionary-based sentiment measures.

Jiang and Zeng (2023) employed ProsusAI's FinBERT model to extract sentiment from large-scale financial news datasets covering more than 800 U.S. companies.[88] Each news headline was processed through FinBERT to obtain probabilities for positive, negative, and neutral sentiment. These probabilities were then aggregated at the daily level for each stock ticker to construct a continuous sentiment score defined as $P(\text{positive}) - P(\text{negative})$.

The daily sentiment scores were combined with lagged values of opening, closing, high, and low prices, as well as trading volume, and were then fed into a LSTM.[89] To assess the incremental contribution of sentiment, the authors compared the FinBERT-enhanced networks against three benchmarks: a LSTM without sentiment input, a canonical ARIMA model, and a LSTM using sentiment derived from the base BERT model.

Their results showed that FinBERT-based sentiment substantially improved predictive performance. While the LSTM alone outperformed the ARIMA baseline, incorporating FinBERT sentiment yielded the strongest forecasts of stock price movements. The gains over the base BERT model further underscored the importance of domain-specific pretraining.[90]

A. H. Huang et al. (2023) introduced a distinct version of FinBERT, hereafter denoted as $FinBERT_2$, which was trained on a different corpa than the model of Araci (2019). The latter is employed in this study and will be discussed in Section **??**.[91] $Finbert_2$ is

[87] The general procedure of finetunging and related topics such as the F1-Score are discussed below.

[88] Several variants of FinBERT exist, this distinction is discussed in more detail below.

[89] A LSTM is a type of recurrent neural network optimized for capturing long-term dependencies across sequential time steps. An intuitive introduction is provided by Olah (2015).

[90] The FinBERT model used in this study was not specifically fine-tuned on financial news. The authors therefore noted that additional task-specific fine-tuning could further enhance predictive accuracy.

[91] Although both models share the same name, A. H. Huang et al. (2023) only briefly mention Araci (2019) in Footnote 43 (p. 834) without clarifying whether their work was inspired by the earlier FinBERT. A preprint of A. H. Huang et al. (2023) had already circulated in 2020 under A. Huang et al. (2020), i.e., shortly after Araci (2019), suggesting that both models may have been developed independently.

also BERT-based model fine-tuned on financial texts, here on regulatory filings, analyst reports, and earnings call transcripts.

Using an annotated dataset of analyst report sentences, the authors benchmarked their model against standard machine learning classifiers as well as several deep learning architectures. Across all evaluation metrics, $FinBERT_2$ achieved the highest accuracy, with particularly pronounced gains in detecting negative sentiment. Notably, unlike the benchmark models, $FinBERT_2$ maintained high accuracy even with limited labeled training data, underscoring the advantages of domain-specific pretraining when annotation is costly.

The model's usefulness is not limited to sentiment. In additional tests, it also performed well in detecting references to environmental, social, and governance issues when being further fine-tuned on those. Applying the model to nearly 29,000 earnings call transcripts, A. H. Huang et al. (2023) further showed that its sentiment measures align more closely with subsequent market reactions than those generated by alternative methods.

To the best of the author's knowledge, no existing study has directly attempted to measure the hype surrounding a specific technology or innovation using Transformer-based models. A key contribution of this thesis will be the construction of such an index - termed the AI Narrative Index (NLP) - based on Araci (2019) FinBERT.

Having established the theoretical foundations of asset pricing (Section 2) and the functioning of FinBERT (Section 3), as well as reviewed the current empirical literature, the following section outlines the exact methodology employed in this study.

5 Methodology

The overall workflow—from news acquisition to inference on financial sentiment—can be divided into three stages: **Phase 1:** Data Collection and Annotation (section 5.1), **Phase 2:** Construction of Narrative Indices (section 5.2, **Phase 3:** Econometric estimation (section 5.3).[92] The subsequent subsections provide a detailed account of each stage.

Final empirical results are reported in Section 6.

5.1 Data Collection and Annotation

Financial data were obtained via the Yahoo Finance API using the `yfinance` library (Aroussi, n.d.) on 01.08.2025. The dataset comprises adjusted stock prices for Apple, Microsoft, Alphabet, Amazon, Meta, NVIDIA, Tesla, Broadcom, AMD, and TSMC, alongside adjusted close prices for the ETFs ROBO, ARKQ, BOTZ, AIQ, and IRBO.[93] The selected companies cover the so-called *Magnificent Seven* and additional semiconductor firms strongly exposed to the current boom in artificial intelligence (Basele et al., 2025). The ETFs were chosen because they not only market themselves as AI-focused, but have also demonstrated strong responsiveness to AI-related news (Ante & Saggu, 2025). Financial data were collected for the period 01 April 2023 to 16 June 2025.

[92] The full codebase is available at https://github.com/LarsIX/narrative_index.

[93] The raw data included the columns "Date", "Ticker", "Open", "High", "Low", "Close", "Adj Close", and "Volume". For subsequent analyses, only "Date", "Adj Close" (the adjusted closing price), and "Ticker" were retained.

Textual data were collected from the digital archives of the WSJ. As briefly mentioned in the introduction, the WSJ was selected for three reasons. First, it is one of the most widely read financial newspaper (Mai & Pukthuanthong, 2021). Second, it has been widely used in prior research to derive measures of investor sentiment (e.g. Mai & Pukthuanthong, 2021; Tetlock, 2007). Third, alternative outlets such as Bloomberg, the Financial Times (FT), Reuters, and CNBC proved technically or legally infeasible to scrape.[94] The corpus comprises all articles published between 01 April 2023 and 16 June 2025 in the sections *politics*, *world*, *finance*, *us-news*, *business*, *opinion*, *tech*, and *economy*, yielding a total of 22,904 articles. The annual distribution was highly imbalanced: 4,401 in 2023, 14,443 in 2024, and 4,060 in 2025. In robustness checks (Section X), daily article counts were included as a control variable but proved insignificant.

All articles were cleaned to remove snippet pollution, hyperlinks, and advertisements. Manual inspection revealed that the vast majority were unrelated to AI, creating a strong class imbalance. To address this, a simple keyword-based dictionary was constructed for pre-filtering before annotation. Twelve terms were identified to ensure broad coverage: *AI, A.I., artificial intelligence, machine learning, deep learning, LLM, GPT, ChatGPT, OpenAI, Transformer model, generative AI, neural network.*[95] According to this filter, for 2024 $\approx 12.61\%$ $(1,821$ out of $14,443$) of articles contained at least one AI reference. Figure 4 below displays the share of flagged versus non-flagged articles by section for 2024. The prominence of AI as a general theme is evident from the fact that roughly 67.44% of all technology-section articles contain at least one fiylter word.

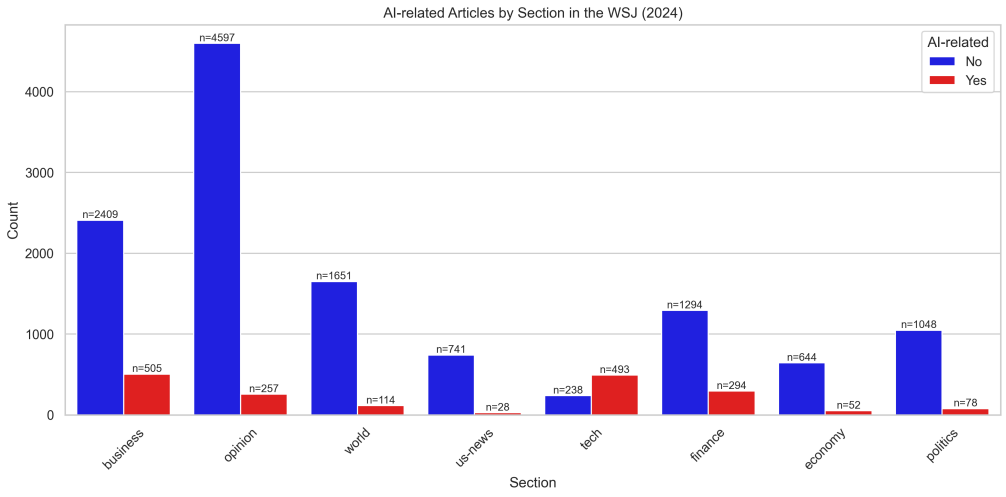

Figure 4. AI-related articles by section in the Wall Street Journal (2024). *Source:* Own illustration.

Manual annotation was conducted on the 2024 subset, which constituted the first dataset available due to computational constraints.[96]

[94]This remained the case even after extensive improvements to the scraping pipeline with GPT-4 assistance. Documentation of prompt iterations is available at: https://chatgpt.com/share/681787c1-c3b8-8013-a542-770e7cf10a84.

[95]Matching was case-insensitive and used word boundaries to avoid partial matches (e.g., "AI" matches "AI" but not "aimless").

[96]Scraping one year of WSJ articles required approximately 5-7 days on the available hardware. As scraping began with 2024, this dataset became available first. Given the time-intensive nature of manual labeling, subsequent years could not be annotated to the same extent.

The first annotation batch was constructed to contain 50% AI-flagged articles based on the keyword filter described above. Following common practice in NLP-research (Gilardi et al., 2023), a second annotator (a paid crowd worker) was recruited to improve generalizability. An initial annotation guide (Appendix III) was drafted and discussed with the annotator. Both reviewers then independently labeled all articles in the batch, using three hype-level categories: (0) unrelated to AI, (1) moderate discussion of AI, and (2) strong emphasis on the potentials of AI.

Disagreement on AI relevance occurred in 19% of cases, and disagreement on hype level in 32%. Most conflicts were resolved through discussion, with final decisions taken by the author.[97] Based on this experience, and following additional literature review, the annotation guide was revised (Appendix III, Table 22 compares the guidelines). Three key adjustments were made: (1) removal of hype-Level 0 for AI-related articles, implying that any discussion for AI was now mapped to a minimum score of 1 (2) expansion of the "moderate hype" definition to include general applications of AI, and (3) stronger emphasis on investor-relevant signals such as stock market impacts and regulatory discussions.

A second batch of 118 articles was then sampled and annotated independently. Of these, 18 were flagged as AI-related by the dictionary. Disagreement on AI relevance occurred in 11.9% of cases. Importantly, whenever both reviewers agreed on AI relevance, they also agreed on the hype level. After conflict resolution, no further revisions to the guide were required. A third batch ($n = 100$) was subsequently annotated, again with 50% flagged articles. Here, disagreements occurred in 8% of cases on AI relevance, and in 15% of hype annotations, corresponding to 7% of all AI-related articles.

Finally, a fourth batch of $n = 700$ articles was annotated by the crowd worker alone. To ensure quality, the author reviewed a 25% random sub-sample, following the validation strategy of S. Baker et al. (2016). In this sub-sample, disagreements on AI relevance increased to 13.1% and disagreements on hype level to 28.6%.[98]

Given the substantial divergence in hype levels - particularly between level 1 and level 2 - the dataset was ultimately used to flag articles as AI-related or not, rather than to assign direct hype scores. In the following analysis, *No Narrative* denotes non-AI-related articles, while *Narrative* refers to AI-related ones.

To extend inference beyond the manually labeled subset, two strategies were pursued: (i) fine-tuning FinBERT, and (ii) applying a dictionary-based classification. Both approaches are discussed in the following section.

To classify the remaining 21,886 articles, a custom FinBERT model based on Araci (2019) was trained.[99]

The 1,018 manually labeled articles were used as ground truth, where the original hype-level 2 was collapsed into 1 to indicate narrative prevalence. The dataset was then

[97] In many cases the final judgment followed the annotator's assessment, underscoring the value of an independent, experienced encoder. Special thanks are extended to Md. Towfique Hossain.

[98] The reason for the higher divergence in this batch remains unclear, though annotation disagreement is a well-documented challenge in the literature (S. Baker et al., 2016; Gilardi et al., 2023).Resource constraints precluded further investigation of this divergence. However, as most inferences relied on dictionary-based filtering (see below), those results are not biased by potential annotation errors.

[99] Preliminary experiments with other BERT-variants such as RoBERTa, DeBERTa, and the original BERT indicated that, across configurations, FinBERT consistently outperformed the alternatives.

split into a stratified 70/20/10 train/validation/test split.[100][101] To improve the model's focus on AI-related text, the flagging mechanism described above was reused to extract relevant snippets. Specifically, input snippets were scanned for filter terms and truncated to a context window of size c. For $c = 0$, only the focal sentence was retained, for $c = 1$, the neighboring sentences were included, and, more generally, for $c = n$, n sentences to the left and right were added.[102]

Further hyperparameters comprised (i) the number of unfrozen layers u, (ii) the dropout rate d, and (iii) the learning rate η (see footnote 67). Here, unfreezing denotes retraining the last u layers, while dropout refers to randomly deactivating a fraction d of neurons in a given layer during training.[103] This technique prevents over-fitting (Srivastava et al., 2014).[104]

Empirical testing of different combinations of c, u, d, and η yielded the best results for $c = 2$, $u = 1$, $d = 0.1$, and $\eta = 2 \cdot 10^{-5}$, with early stopping after 5 epochs.[105]

The final model achieved strong performance on the validation set with an overall accuracy of 0.92. Both the macro and weighted F1-scores reached 0.92, while the class-specific F1-scores were 0.93 for the *No Narrative* class and 0.91 for the *Narrative* class.[106]

As a benchmark, and inspired by the dictionary-based approach of S. Baker et al. (2016) in constructing the EPU, a dictionary was developed to label AI-related content. Three publicly available AI glossaries were used as sources: by the International Monetary Fund (Fund, 2023), Stanford University (Manning, 2020), and the Parliament of the United Kingdom (Gajjar, 2024). Together those yielded 49 terms.[107]

After validating occurrences in a subset of articles, the terms *Bias*, *Compute*, *Disinformation*, *Educational Technology*, and *Robotics* were excluded, as they frequently appeared in unrelated contexts.

[100] Stratification preserved class balance across all subsets.

[101] In machine learning, the available data are typically split into separate subsets before training begins. The training set is used to fit the model, while the validation set is used to monitor intermediate performance and to check whether the model generalizes beyond the training data. Final evaluation is performed *once* on the test set to assess true out-of-sample performance.

[102] During initial experiments, it became evident that with such a small annotated dataset, fine-tuning FinBERT without any additional mechanism to flag AI-related articles led to poor performance - the model was essentially unable to distinguish between AI- and non-AI-related content.

[103] Random deactivation is performed epoch-wise.

[104] Overfitting refers to a situation where the model has memorized the training data 'too good', including its noise, making it perform badly on unseen data.

[105] An *epoch* refers to one full pass of the entire training set. *Early stopping* prevents over-fitting by stopping the training when the model stops improving on the test data.

[106] Accuracy is defined as the proportion of correctly classified samples. Precision and recall are given by Precision $= \frac{TP}{TP+FP}$ and Recall $= \frac{TP}{TP+FN}$, where TP, FP, and FN denote true positives, false positives, and false negatives. The F1-score is their harmonic mean, $F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$. Macro F1 averages F1 equally across classes, while weighted F1 weights them by the relative number of classes in the data (Terven et al., 2023).

[107] I.e. Large Language Model, Large Language Models, LLM, Bias, Machine Learning, Deep Learning, Neural Network, Fine-Tuning, Prompt Engineering, Generative AI, Prompts, Hallucination, Supervised Learning, Unsupervised Learning, Autonomous Systems, Reinforcement Learning, Algorithm, Narrow AI, Speech Recognition, Facial Recognition, Human-level AI, Artificial General Intelligence, AGI, Social Chatbots, Human-Robot Interaction, Human-Centered Artificial Intelligence, Algorithmic Bias, Artificial Neural Network, Automated Decision-Making, Compute, Computer Vision, Deepfake, Deepfakes, Disinformation, Educational Technology, Foundation Model, Foundation Models, Frontier AI, Generative Adversarial Networks, Graphical Processing Units, Interpretability, Open-Source, Responsible AI, Robotics, Superintelligence, Training Datasets, Transformers, Transformer Artificial Intelligence.

For reference, the dictionary based on this final set will be referred to as the *naive label dictionary*, while the dictionary used by the custom FinBERT model will be denoted as the *flagging dictionary*.

The naive label dictionary achieved an accuracy of approximately 87.74% when applied to the manually labeled dataset. A direct comparison with predictions from the custom FinBERT model yielded an agreement rate of 97.9%. This highlights the effectiveness of relatively simple dictionary-based methods for identifying AI-related articles, though it should be noted that FinBERT was explicitly trained using a smaller dictionary as input features.

To examine this further, Table 24 in Appendix III compares the dictionaries. Notably, only six of the twelve terms in the smaller flagging dictionary also appear in the naive-labeling dictionary. For brevity, only the most striking points are discussed here: the terms *AI*, *A.I.*, and *Artificial Intelligence* were absent from the naive dictionary, as were *GPT*, *ChatGPT*, and *OpenAI*.[108] The intersection consists of *Machine Learning*, *Deep Learning*, *Neural Network*, *LLM*, *Generative AI*, and *Transformer*. Given this relatively limited overlap (50% of the naive dictionary terms do not appear in the flagging dictionary), the high level of agreement is remarkable.

To further investigate this issue, Table 24 reports the coverage of terms across the naive dictionary, the extended flagging dictionary, and their intersection.

These raw labels were subsequently used to infer sentiment via the FinBERT model of Araci (2019), notably without further fine-tuning.[109] The results are discussed in the following section.

5.2 Construction of Narrative Indices

Two distinct strategies were employed to infer sentiment. First, only those articles labeled as AI-related by either the fine-tuned FinBERT model (2,941 of 21,886, $\approx$ 13.7%) or by manual annotation (405 of 1,018, 39.8%) were fed into the pretrained FinBERT sentiment model.[111] To address the 512-token limit (see above), the headline, sub-headline (if available), and article body were concatenated and subsequently truncated, prioritizing earlier tokens. This procedure ensured that headlines were always fully included,[112] as they are known to be highly informative (Nassirtoussi et al., 2015).[113] The model's default sentiment categories (negative, neutral, positive) were employed, as this setting has been shown to produce robust results (Acheampong et al., 2021).

Second, after discarding the AI-related labels, all 22,904 articles were reprocessed with FinBERT, this time using the naive label dictionary for snippet extraction. For each

[108] One might argue that including *AI*, *A.I.*, and *Artificial Intelligence* in the naive dictionary would have been appropriate. However, the rationale was to construct one hand-crafted dictionary based on manual inspection of a subset of WSJ articles, and one "official" dictionary derived from established glossaries. For the latter, only polluting key-words were deleted and none added.

[109] This refers to the original [110] sentiment model, not the fine-tuned variant introduced above.

[111] Model available at https://huggingface.co/ProsusAI/finbert.

[112] As a rule of thumb, 512 tokens correspond to approximately 256 words (Sennrich et al., 2015).

[113] To the best of the author's knowledge, no systematic benchmark of truncation techniques for news-based sentiment inference exists. While further study of this issue would be valuable, it lies beyond the scope of this thesis.

article, the headline, sub-headline (if available), and corpus were concatenated. The concatenated text was then scanned for occurrences of dictionary terms, and for each match, the sentence of occurrence together with c neighboring sentences was extracted as the context window. For inference, the same default classification (negative, neutral, positive) was employed. Initial experiments indicated that context windows larger than $c = 2$ under-performed in subsequent regressions (see below). Accordingly, results for $c \in \{0, 1, 2\}$ were retained for further analysis.

Next, sentiment categories were mapped to numerical scores (negative $= -1$, neutral $= 0$, positive $= 1$) following common practice (Acheampong et al., 2021). Summing these scores on a daily basis produced four narrative indices spanning 781 days, denoted as $AINI$:

- **AINI**$_{\text{custom}}$: Based on articles labeled as AI-related by the fine-tuned FinBERT model and manual annotation, inferred on the full corpus (subject to the truncation limit).
- **AINI**$_{w_0}$: Based on the naive label dictionary with a context window of 0.
- **AINI**$_{w_1}$: Based on the naive label dictionary with a context window of 1.
- **AINI**$_{w_2}$: Based on the naive label dictionary with a context window of 2.

For comparability, each daily raw index was normalized by dividing by the total number of articles published on that day, yielding $AINI_t^{\text{norm}}$. Building on this normalized index, two additional Exponential Moving Average (EMA)s with different smoothing parameters were constructed:

$$EMA_t = \alpha\, AINI_t^{\text{norm}} + (1 - \alpha)\, EMA_{t-1}, \quad \text{with } \alpha \in \{0.2,\, 0.8\}. \tag{19}$$

The terms $EMA_{0.2,t}$ and $EMA_{0.8,t}$ denote exponential moving averages of the normalized index, where α is the smoothing factor that determines the relative forgetting curve. A lower α (e.g., 0.2) results in stronger smoothing, while a higher α (e.g., 0.8) results in faster forgetting. Before analyzing various time series, the next subsection will discuss empirical estimation.

5.3 Econometric estimation

All time series were tested for stationarity using three standard tests: Augmented Dickey–Fuller Test (ADF), Phillips–Perron Test (PP), and Kwiatkowski–Phillips–Schmidt–Shin Test (KPSS) tests.[114]

As shown in Table **??** in Appendix I, where for the AINI variants the subscripts indicates the model utilized (c=*custom*, wk = naive label dictionary with context size k) all

[114]The ADF test examines whether an autoregressive process is non-stationary, i.e., whether it contains a unit root (Dickey & Fuller, 1979). The PP test adjusts for serial correlation and heteroskedasticity in the error term (Phillips & Perron, 1988). In contrast, the KPSS test reverses the null hypothesis by testing for stationarity (Kwiatkowski et al., 1992).

financial and control variables are stationary across all periods and test specifications, apart from the logged growth of daily number of articles in 2025. For the AINI measures, however, results are more nuanced. In particular, the 2025 period exhibits several cases of non-stationarity. The $EMA_{0.2}$ measure is consistently identified as non-stationary by all three tests across all model variants and is therefore excluded from subsequent analysis. Similarly, $AINI^{norm}$ is rejected by both the KPSS and ADF tests in 2025, motivating the exclusion for that period as well.

Interestingly, all $AINI_c$ (i.e. custom) measures remain stationary in 2025, while none of the window-based variants ($w0$-$w2$) do. This asymmetry suggests potential structural change or narrative accumulation during that period - possibly reflecting stronger persistence in AI-related discourse. Since standard Granger causality models require stationarity, the non-stationary window-based measures for 2025 are excluded from further tests, though their behavior is itself economically informative.

When extending the analysis to the periods 2023-2024 and 2024-2025, the picture becomes more mixed. In these sub-samples, the ADF and PP tests generally reject the unit-root hypothesis, implying stationarity , whereas the KPSS test rejects stationarity for nearly all *custom* models (Dickey & Fuller, 1979; Kwiatkowski et al., 1992; Phillips & Perron, 1988), thus making the here-applied Granger causality tests unsuitable (Dolado & Lütkepohl, 1996). As none of the *custom* models yielded statistically significant results at any point (see next section), this issue is not examined further. Consequently, the case of the KPSS test rejecting non-stationarity for logged article growth in 2025 - while itself an interesting observation - does not pose an obstacle, as *custom* was the only model evaluated for that period.

Next, to investigate linkages between stock prices and the constructed indices, a series of Granger causality tests were conducted. Formally, Granger causality evaluates whether lagged realizations of a variable X_t contain additional predictive information for another variable Y_t beyond that captured by its own lag structure Y_{t-i}, $i = 1, \ldots, l$.

This framework aligns quite well with the logic of the EMH discussed in Section 2.1: remember that under the semi-strong (strong) form, current prices P_t should fully incorporate all publicly (privately) available information (Fama, 1991; Fama & French, 1993; Malkiel, 2003). Consequently, once lagged prices (or here, returns) are controlled for, any statistically significant predictive content of an additional variable would either indicate market inefficiency or - as discussed in Section 2.2 - reflect sensitivity to changing risk factors.

Daily log returns were regressed on their own lagged values. Using logarithmic returns not only promotes stationarity and ensures symmetry in price changes (a doubling and a halving have the same magnitude on the number line), but also allows interpreting these changes approximately as percentage variations. This is a standard transformation in empirical finance and asset-pricing research (**tsayAnalysisFinancialTime2010**; see, e.g., Merton, 1980).[115]

In a second specification, one of the three AINI variants was included as an additional regressor, and a Wald test was employed to assess whether its inclusion significantly

[115]Mathematically, R_t denoting the simple return given by first differences, logged returns are defined as $r_t = \ln(P_t/P_{t-1}) = \ln(1 + R_t)$, ln denotes the natural logarithm. By a first-order Taylor expansion of $\ln(1 + R_t)$ around $R_t = 0$ (which approximately holds in our case, see Table 20), one obtains $r_t \approx R_t = (P_t - P_{t-1})/P_{t-1}$. Thus $r_t = 0.01$ roughly equals a 1% price increase (Tsay, 2005, pp. 5–6).

enhanced predictive power.[116]

In line with established approaches regressing market movements on news data or social media data (e.g. Gupta & Chen, 2020; Tetlock, 2007), only trading days - excluding weekends and banking holidays - were considered. Corresponding AINI observations for non-trading days were thus omitted.[117]

To test the robustness of results, three additional covariates were included to control for underlying risk factors: the S&P 500 index (proxying the CAPM market *beta*), the VIX (capturing aggregate market risk), and the SOX index (accounting for industry-wide dynamics in semiconductors). To control for potential bias due to the variance in daily articles, a fourth test was run with the number of daily article counts as a control variable.

All control variables enter the regression as log growth rates, computed as the first differences of the logarithmic closing values of consecutive trading days.[118][119] Heteroskedasticity was addressed using Newey-West heteroskedasticity- and auto-correlation-consistent standard errors (Newey & West, 1987).

Mathematically, the following regression model was estimated:

$$R_t = \alpha + \sum_{\ell=1}^{l} \gamma_\ell \, AINI_{t-\ell} + \sum_{\ell=1}^{l} \beta_\ell \, R_{t-\ell} + \sum_{m=1}^{M} \sum_{\ell=1}^{p_{cm}} \zeta_{m,\ell} \, C_{m,t-\ell} + u_t, \tag{20}$$

$$H_0 : \gamma_1 = \cdots = \gamma_{p_x} = 0$$

Here, $AINI_t^{\text{var}}$ denotes one of the three variants of the AI Narrative Index, R_t is the daily log return of the respective stock or Exchange-Traded Fund (ETF), and C_t represents the control variable. Accordingly γ denotes the regression coefficient on $AINI_t^{\text{var}}$, β on on log returns and ζ on the given control variable.

The null hypothesis H_0 states that the AINI contains no incremental predictive information for future returns beyond that captured by past returns. The parameter α is a constant term, u_t is the error term, and l denotes the common lag length.

As three measures of the AINI were constructed for each of the four models, and for five ETFs plus ten individual stocks, this yields $3 \times 4 \times 15 = 180$ Granger causality regressions per time period.

To capture potential time variation, these tests were repeated for six time periods: 2023 (188 trading days), 2024 (252 trading days), 2025 (132 trading days), 2023–2024 (440 trading days), 2024–2025 (384 trading days), and the full period 2023–2025 (572 trading days). This resulted in $180 \times 6 = 1{,}080$ Granger causality tests for a given lag length l. Empirically, significance declined sharply for $l > 3$; thus, $l \in \{1, 2, 3\}$ was chosen. Consistent with prior studies (e.g. Bollen et al., 2011; Jansen & Nahuis, 2003; Tetlock, 2007),

[116] Intuitively, we can think of the Wald test as testing whether the inclusion of a given AINI variable reduced prediction error compared to the past-returns only specification. See footnote 142 for further explanation.

[117] This procedure introduces the noteworthy feature that the EMA-based variants may embed information not fully captured by the $AINI^{norm}$.

[118] For brevity, we refer to the S&P 500, VIX, and related measures by name. As outlined above, these represent logged growth rates.

[119] Significance of these controls would itself be noteworthy, as it would suggest that AI narratives co-move with broader or sector-specific market conditions.

lag lengths were held constant across variables, leading to $1{,}080 \times 3 = 3{,}240$ regression pairs. Excluding the 405 non-stationary time series discussed above reduced the sample to 2,835 valid regression pairs.[120]

To test for reverse Granger Causality - whether rising stock prices might themselves trigger increased AI-related narratives - the regressions were also estimated in the opposite direction, with $AINI^{var}$ as the dependent variable and log returns as predictors, yielding $2{,}835 \times 2 = 5{,}670$ total regressions. Using the same notations as above, the model can be expressed as:

$$AINI_t = \alpha + \sum_{\ell=1}^{l} \beta_\ell \, R_{t-\ell} + \sum_{\ell=1}^{l} \gamma_\ell \, AINI_{t-\ell} + \sum_{m=1}^{M} \sum_{\ell=1}^{p_{cm}} \zeta_{m,\ell} \, C_{m,t-\ell} + u_t, \tag{21}$$
$$H_0 : \beta_1 = \cdots = \beta_l = 0$$

While using the same denotations as above, rejecting the null hypothesis $H_0^{(AINI \nRightarrow R)}$ implies here that past values of $AINI$ contain statistically significant predictive information for returns beyond their own auto-regressive dynamics.

Including models both with and without controls, a total of $5{,}670 \times 4 = 22{,}680$ Granger causality tests were conducted. Full regression outputs are provided in the online appendix,[121] while the most relevant results are summarized in table **??**.

A final concern to be addressed is the potential occurrence of *collider bias*. This arises when two otherwise unrelated variables both exert an effect on a third, so that conditioning on the latter opens a spurious backdoor path and induces correlation between the former (see Elwert & Winship, 2014; Pearl, 2009) and subsection 5.3.

In our context, the potentially problematic collider is the VIX. First, market movements are known to trigger changes in the VIX (Pathak & Deb, 2020);[122] second, as discussed in subsection 5.3, including the VIX substantially increases the number of rejected regression pairs. If both returns and AINI influence the VIX, conditioning on VIX could therefore introduce a spurious correlation between AINI and returns rather than eliminating one.

To test for this possibility, the Granger-causality framework was again. Here values of the VIX were regressed on their own lags and, in the unrestricted model, additionally on lagged AINI terms:

$$VIX_t = \alpha + \sum_{\ell=1}^{p_x} \gamma_\ell \, AINI_{t-\ell} + \sum_{m=1}^{M} \sum_{\ell=1}^{p_{cm}} \zeta_{m,\ell} \, C_{m,t-\ell} + u_t, \tag{22}$$
$$H_0 : \gamma_1 = \cdots = \gamma_{p_x} = 0 \ .$$

Rejection of H_0 in this setup would imply that AINI Granger-causes the VIX, consistent with a potential collider relationship: both returns and AINI could influence VIX, such

[120] This reduction corresponds to 15 assets $\times 3$ models $\times$ $3 AINI^{var}$ $\times 3$ lags $= 405$ regression pairs for the window-based models in 2025, plus an additional 45 pairs excluded for the $EMA_{0.2}$ variant in the custom model for 2025.

[121] https://github.com/LarsIX/narrative_index/tree/main/reports/tables

[122] Empirically, declines in stock prices can increase the VIX. According to the leverage hypothesis, falling equity values raise firms' debt-to-equity ratios, which causes investors to perceive the asset as riskier, which increase trading patterns and causes an increase in the VIX (Pathak & Deb, 2020).

that controlling for VIX in return regressions might induce dependence between otherwise distinct processes.[123]

Turning to the general framework, empirical p-values for all regressions were obtained using a *null-imposing bootstrap* with 10,000 replications, following Kónya (2006).

Specifically, the procedure combined the *studentized restricted-model bootstrap* of **LiMaddala1996<em** T_2/S_3 variant) with the *restricted wild bootstrap* of Davidson and MacKinnon (2010), also utilized by Davidson and MacKinnon (2010), adapted here to the Granger-causality regression in (20).[124]

Formally, let R_t again denote the daily log return of a stock or ETF, $AINI_t^{\text{var}}$ one of the variants of the AI Narrative Index ($AINI^{\text{norm}}$, $EMA_{0.2}$, $EMA_{0.8}$), C_t control variable (in log-differenced form) and $H_0 : \gamma_1 = \cdots = \gamma_l = 0$ be the null hypothesis, then the bootstrapped algorithm proceeded as follows (see https://github.com/LarsIX/narrative_index/blob/main/src/modelling/estimate_granger_causality.py):

1. **Restricted estimation under H_0:** Estimate the restricted model that excludes lagged $AINI^{\text{var}}$ terms from (20), obtaining restricted fitted values $\hat{R}_t^{(0)}$ and residuals $\hat{u}_t^{(0)}$.

2. **Wild residual generation:** Draw i.i.d. wild bootstrap weights w_t with $\mathbb{E}[w_t] = 0$ and $\text{Var}(w_t) = 1$, from the Rademacher distribution ($P(w_t = \pm 1) = 0.5$), and form bootstrap residuals $u_t^* = w_t \cdot \hat{u}_t^{(0)}$.

3. **Bootstrap sample generation:** Construct the bootstrap-dependent variable $R_t^* = \hat{R}_t^{(0)} + u_t^*$, which enforces the null-imposed conditional mean while preserving the (potentially heteroskedastic) error structure.

4. **Unrestricted re-estimation:** Estimate the unrestricted regression model on the bootstrap sample $(R_t^*, AINI_{t-\ell}^{\text{var}}, R_{t-\ell}, C_{m,t-\ell})$ and compute covariance estimators.

5. **Bootstrap p-value:** Repeat steps 2-4 for $b = 1, \ldots, B$ bootstrap replications. Let F_{obs} denote the observed Wald statistic from the unrestricted model. The empirical (bootstrapped) p-value is then estimated by

$$\hat{p}_{\text{boot}} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}(F_b^* \geq F_{\text{obs}})}{B + 1},$$

   where $\mathbb{I}(\cdot)$ denotes the indicator function.

This procedure ensured that bootstrapped samples: (i) satisfy the null, (ii) preserve the original regressor matrix, and (iii) replicate heteroskedasticity. The same procedure was, in principle, also applied to the reverse regressions in (21), where the null hypothesis $H_0 : \beta_1 = \cdots = \beta_l = 0$ tests whether past returns Granger-cause the given AINI index. It was likewise applied to the test of a potential collider relationship between AINI and returns through the VIX (see equation (22)).

[123] As detailed below, no regression pair exhibited statistical significance. Consequently, an additional regression of logged returns on the VIX was not warranted.

[124] Unlike the canonical bootstrap, the wild bootstrap preserves the heteroskedastic structure of residuals by multiplying each one with a random weight ('wild' shocks) rather than resampling.

To address multiple testing and improve empirical rigor, the p-vales for each conducted test were adjusted using the Benjamini–Hochberg Method (BH) procedure (Benjamini & Hochberg, 1995). The algorithm proceeded as follows (see https://github.com/LarsIX/narrative_index/blob/main/src/modelling/estimate_granger_causality.py):

1. **Define testing families:** Within each *ticker-period-direction* combination g, collect the set of hypotheses associated with the three AINI variants:

$$\{AINI^{\text{norm}},\ EMA_{0.2},\ EMA_{0.8}\}.$$

   Denote the corresponding p-values as $\{p_{g,v}\}_{v=1}^{V_g}$, where $V_g = 3$.

2. **Order p-values:** Sort the p-values in ascending order within each family:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(V_g)}.$$

3. **Compute BH critical values:** For each $p_{(i)}$, compute the Benjamini-Hochberg threshold

$$p^*_{(i)} = \frac{i}{V_g}\, q, \quad \text{with } q = 0.05.$$

4. **Determine significant results:** Find the largest index k such that

$$p_{(k)} \leq p^*_{(k)}.$$

   Reject all null hypotheses $H_{0,(i)}$ for $i \leq k$.

5. **Apply to analytic p-values:** The same procedure was applied to the analytic Wald F p-values $p^{\text{ana}}_{g,v}$.

This procedure ensured that, within each family, the expected proportion of false reections remained below $q = 0.05$, thereby limiting Type I error (falsely rejecting H_0, often named false positive) inflation (Benjamini & Hochberg, 1995).

The next section will discuss the empirical results, by first giving a descriptive overview and then diving deeper into regression results in context of the theoretical background given in sections 2.1 and 2.3.

6 Results

6.1 AI Narrative Indices

Before analyzing the time series directly, it is helpful to first develop an intuition for how the underlying sentiment scores were generated. This can be done by dividing the articles used for inference into three broad categories: (i) those primarily about AI, (ii) those that mention AI as a secondary or side topic, and (iii) those unrelated to AI (false positives).

One representative example of each category is presented in Appendix II, Subsection B.1, and it is useful to briefly discuss their characteristics.

Example 1 is genuinely AI-related and reports how ASML has benefited from trade tensions between Western economies and China. Capturing only the phrase that ASML holds a monopoly (w_0) led FinBERT to infer a neutral sentiment. Expanding the context to w_1 and w_2 added information about the firm's valuation - "its stock, at 33 times forward earnings, trades at a premium to peers" -which shifted sentiment to positive. However, the (near) full-article context (*custom*) reversed the sign again as geopolitical risks and trade-war frictions became salient. Hence, w_1 and w_2 may be viewed as picking up genuine optimism about AI-related profits, while the *custom* model captured the broader negative narrative of geopolitical tension. The example thus illustrates how varying context length alters which aspect of a story - financial opportunity or political risk - dominates the model's sentiment assessment.[125]

Example 2 discusses how the so-called "Proud Boys"[126] prepare for the 2024 presidential election and the potential threat they pose to democratic stability. Within this article, AI plays only a minor role: it is briefly mentioned in the context of Telegram - the group's main communication platform - using AI to monitor hate speech. The w_0 and w_1 windows capture Telegram's statement that it will "keep the platform safe for users" through AI, which FinBERT scores positively ($+1$). The w_2 window yields a neutral score, likely because it also includes passages about Telegram's opposition to violence and disruption of democratic processes - topics with negative emotional valence. When the full article context (*custom*) is included, the sentiment turns negative (-1) as discussions about political extremism and potential unrest dominate the narrative. In sum, these examples illustrate how small context windows can isolate and capture AI-related narratives even when the broader article concerns an entirely different topic, whereas larger windows increasingly absorb the surrounding, often unrelated, narrative context.

Example 3 illustrates a clear false positive: the article is unrelated to AI but was selected because of the word "prompt." Nevertheless, it reveals the mechanics of FinBERT's sentiment inference when comparing w_0 and w_1. While "energy's capex commitments prompt" (w_0) yielded a score of $+1$, extending the window to "Woodside Energy's capex commitments prompt Macquarie to downgrade the stock to neutral from outperform" (w_1) flipped the score to -1. FinBERT likely associated "capex commitments" with positive corporate investment, but revised its interpretation once the context indicated a downgrade.

Turning to the time-series constructed from these models, valuable initial insights can be obtained by averaging across all three variants ($AINI^{\text{norm}}$, $EMA_{0.2}$, and $EMA_{0.8}$) for each model specification.[127] The resulting time series, displayed in Figure 5, are plotted alongside major AI-related events to visualize their temporal co-movements, revealing several noteworthy patterns.[128]

[125]The exact rationale behind FinBERT's inferences cannot be determined with certainty. However, given its training data and documented benchmark performance, the qualitative interpretations discussed above are plausible (Araci, 2019; Devlin et al., 2019).

[126]A group of far-right supporters of Donald Trump who played a major role in the Capitol riots following his electoral defeat to Joe Biden in 2020 (Wendling, 2023).

[127]Raw values for non-stationary $AINI$ time-series were included in descriptive analysis for the whole period, as those are only non-stationary in 2025.

[128]The selection of major events reflects the author's informed judgment, given the absence of an established or universally accepted AI timeline.

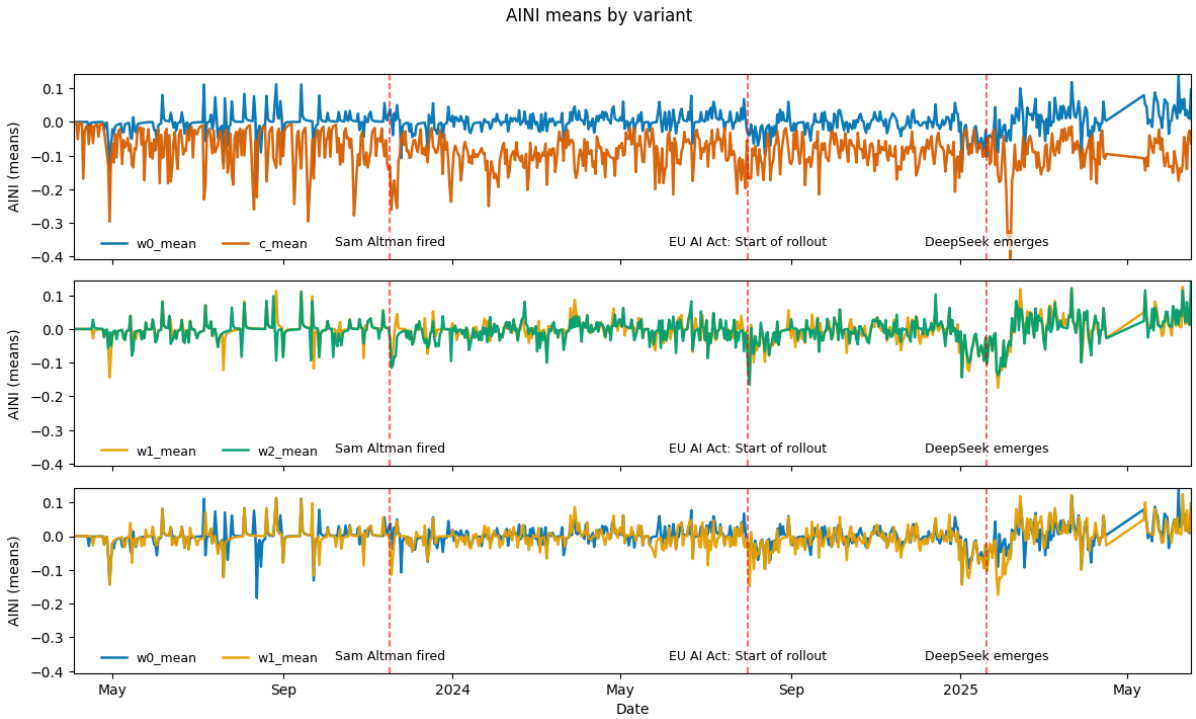

Figure 5. Average $AINI$ values across variants by model. *Source:* Own illustration.

First, although distinct from each other, all naive-dictionary-based variants diverge significantly from $AINI_{custom}$, as the latter exhibits higher variance and, during certain periods, shows markedly different dynamics. For example, in April and May 2025, all window-based measures increased, whereas the custom measure decreased.

Second, while all indices exhibit negative spikes around major events - such as the dismissal of OpenAI chairman Sam Altman,[129] the adoption of the EU AI Act,[130] and the emergence of DeepSeek,[131] - $AINI_{custom}$ displays additional spikes not clearly tied to these events. A similar yet less pronounced pattern emerges in the window-based indices, particularly until October 2023 and again following the downturn after the DeepSeek release.[132]

Third, visual inspection suggests that the Altman dismissal and the EU AI Act were followed by immediate drops in the indices, whereas the effect of DeepSeek appears more delayed, possibly because public recognition of its significance took longer to develop.

Finally, manual inspection indicates that only a small fraction of articles during these three spikes reported directly on the events themselves. Most instead discussed broader

[129]In November 2023, OpenAI's board unexpectedly removed CEO Sam Altman, causing widespread uncertainty and rumors about the future direction of the company. After discussions with Microsoft (OpenAIs main investor) he was re-installed (Hagey, 2025).

[130]The EU AI Act, provisionally agreed in December 2023, introduced one of the world's strictest regulatory frameworks for AI (Neuwirth, 2023) and entered into force step-wise following 01.08.2024 (Commission, 2024).

[131]DeepSeek, a Chinese-developed AI model released in mid-2024, gained attention as a potential competitor to leading Western foundation model. Crucially, it depends on far less computational power than competing models (Wirtschaftswoche, 2025).

[132]The divergence between the window-based models and $AINI_{custom}$ in 2025 - especially around May - might also account for the differing outcomes observed in the stationarity tests.

AI-related topics. This suggests that such events might have acted as catalysts for heightened overall attention to AI, rather than being the sole subjects of reporting.

Next, period-wise averages and standard deviations of $AINI^{norm}$ are examined for each model. Single-year data for 2025 are excluded due to non-stationarity (see above).[133] Let c denote the custom model and w_i the window-based models of size i as above, with standard deviations shown in parentheses, we obtain:

Year / Model	w0_mean (std)	w1_mean (std)	w2_mean (std)	c_mean (std)
2023	-0.0021 (0.0479)	-0.0026 (0.0414)	-0.0060 (0.0425)	-0.0749 (0.0878)
2024	0.0005 (0.0321)	-0.0055 (0.0393)	-0.0055 (0.0408)	-0.0937 (0.0598)
2023–2024	-0.0006 (0.0396)	-0.0043 (0.0402)	-0.0057 (0.0415)	-0.0857 (0.0735)
2024–2025	0.0012 (0.0435)	-0.0040 (0.0503)	-0.0039 (0.0504)	-0.0943 (0.0652)
2023–2025	0.0001 (0.0451)	-0.0035 (0.0473)	-0.0047 (0.0478)	-0.0876 (0.0744)

Table 1. Mean and standard deviation of $AINI^{norm}$ per model. *Source:* Own illustration.

Across all periods, mean values remain close to zero.[134] Standard deviations, are substantially larger than the means, particularly for the window-based models: for w_0–w_2 they range between 0.03 and 0.05, while for the *custom* measure - whose means are roughly an order of magnitude larger - they range between 0.06 and 0.09. This indicates considerable short-term dispersion relative to the low long-run averages, suggesting that daily sentiment fluctuates widely even when its central tendency remains near zero. Positive mean values occur only for w_0 in 2024, 2024–2025, and 2023–2025, while all other means are negative. The relative ordering of measures is stable across periods: w_0 remains the most positive, followed by w_1, w_2, and finally the *custom* model, which exhibits the lowest values throughout.

From left to right, comparing means across periods for the same period, absolute magnitudes tend to increase.[135] For w_0 and w_2, values in 2024 are slightly higher (less negative) than in 2023, whereas w_1 and *custom* display the opposite pattern. The overlapping periods 2023–2024 and 2024–2025 fall between their adjacent single-year averages, while standard deviations remain broadly stable over time, indicating that overall dispersion does not vary materially across years. The *custom* measure, which captures near full-article context up to the 512-token limit,[136]

To investigate these relationships more systematically, ratios of mean values between models are calculated. The results are shown in Table 2.

[133] For non-stationary time series, the sample mean ceases to be the maximum-likelihood estimate and thus loses its intuitive interpretability (PŸUR, 2018).

[134] By construction, individual article scores are bounded between -1 and 1. Aggregation and daily normalization further compress this scale, so small mean values do not imply irrelevance. The remaining variation is still sufficient for statistical inference.

[135] An exception occurs in 2024–2025, where w_2 is marginally less negative than w_1 (by 0.0001).

[136] For simplicity, this is referred to as (near) full-article context below.

Year / Window	c/w_2	c/w_1	c/w_0	w_2/w_1	w_2/w_0	w_1/w_0
2023	12.5436	28.5539	36.0448	2.2764	2.8736	1.2623
2024	16.9847	17.0738	-182.3434	1.0052	-10.7357	-10.6797
2023–2024	15.0132	20.0661	146.1982	1.3366	9.7380	7.2858
2024–2025	23.8918	23.3596	-75.5682	0.9777	-3.1629	-3.2350
2023–2025	18.8301	24.6943	-945.0509	1.3114	-50.1882	-38.2700

Table 2. Ratios of means of $AINI^{norm}$ between models. *Source:* Own illustration.

Due to near-zero means of w_0 in 2024 and 2023–2025, ratios involving this measure become numerically unstable (Franz, 2007). These cases are shaded grey and should be disregarded. Nevertheless, several patterns emerge. From left to right, ratios generally increase in absolute magnitude as the contextual gap between numerator and denominator widens, confirming that broader textual context amplifies negativity. For 2023, values such as $w_1/w_0 = 1.26$, $w_2/w_0 = 2.87$, and $c/w_0 = 36.04$ show a monotonic rise as additional context is incorporated. This suggests that isolated AI-related sentences are weakly polarized, whereas their surrounding paragraphs substantially deepen the tone. The jump from w_2 to *custom* ($c/w_2 = 12.54$) illustrates how sentiment becomes non-linearly more negative when near full-article context is introduced. For most periods, w_2/w_1 ratios remain close to one, while all valid ratios involving w_0 are relatively large (up to approximately 146 for c/w_0 in 2023–2024).

Taken together, these findings reveal a structural asymmetry between local and extended context in financial news sentiment. AI-related statements appear comparatively neutral in isolation, yet when embedded within full-article narratives, the overall tone consistently turns negative. For empirical analysis, this cautions against using short-text sentiment as a proxy for broader market narratives, as it may systematically understate prevailing tone. Conversely, excessively broad textual windows risk conflating topic-specific sentiment with unrelated contextual information, thereby diluting signal relevance. The result highlights the importance of aligning textual scope with the underlying economic reference frame when constructing sentiment-based indicators.

6.2 Assets

To analyze the characteristics of the assets under investigation, their behavior is examined within the canonical CAPM framework through the parameters α and β. Following Fama and French (1992), daily logarithmic excess returns are regressed on logarithmic contemporaneous market excess returns according to

$$\ln(1 + r_{i,t}) - \ln(1 + r_{f,t}) = \alpha_i + \beta_i \Big[\ln(1 + r_{m,t}) - \ln(1 + r_{f,t})\Big] + \varepsilon_t \tag{23}$$

where $r_{i,t}$ denotes the return on asset i, $r_{m,t}$ the market return (S&P 500), and $r_{f,t}$ the risk-free rate, proxied by the three-month U.S. Treasury bill yield converted from its

annualized percentage form to a daily equivalent. Data were again retrieved from the Yahoo Finance API via the `yfinance` library (Aroussi, n.d.).[137]

In this specification, α_i captures the average abnormal daily *log* excess return (in %), while β_i measures the sensitivity of asset i to market-wide excess returns. To facilitate interpretation, the reported α_i values were transformed into total compounded abnormal returns over the full sample of $T = 572$ trading days according to

$$R_{\alpha,\text{total}} = \exp(\alpha_i T) - 1 \tag{24}$$

which expresses the cumulative excess return implied by α_i under continuous compounding (**arratiaComputationalFinance2014**) across all trading days in the sample.

While the complete regression output is provided in Table **??** in Appendix **??**, the table below summarizes the key results by directly comparing R^2, β, and continuously compounded α coefficients alongside cumulative performance over the same period. The latter was computed by rebasing all assets to 100 on 3 April 2023 and used to sort the table. The dashed line marks the threshold for outperforming the market, which is proxied by the S&P 500 index, and ETFs are shaded in grey.

Table 3. Total Growth, Compounded α, β, and R^2 of Selected Stocks (Sorted by Total Growth)

Ticker	**Total Growth (%)**	$\boldsymbol{\alpha}$ (total, %, compounded)	$\boldsymbol{\beta}$	$\boldsymbol{R^2}$ (%)
ROBO	10.82	-31.36	1.190	75.80
BOTZ	27.73	-24.25	1.333	74.90
AAPL	27.28	-20.41	1.191	49.30
IRBO	34.11	-23.25	1.444	74.20
S&P500	53.38	–	1.000	100.00
TSLA	59.56	-34.45	2.325	34.50
AMD	61.15	-24.44	2.053	41.40
GOOGL	75.44	11.83	1.124	35.10
MSFT	79.23	16.06	1.041	48.90
AIQ	79.58	6.55	1.323	85.30
ARKQ	89.17	2.92	1.572	69.50
META	235.24	89.69	1.508	43.40
TSM	164.49	45.22	1.582	39.90
AMZN	121.02	27.14	1.409	49.50
AVGO	352.44	118.97	2.003	40.90
NVDA	510.84	181.28	2.182	42.70

Source. Own calculations, data derived from Yahoo Finance at 10.10.2025. Prepared with the assistance of generative AI (https://chatgpt.com/share/68df816e-d7b4-8013-874b-ccd146b5823b).

[137] Estimation was conducted using OLS with Newey–West heteroskedasticity- and autocorrelation-consistent standard errors (Newey & West, 1987).

As expected, all assets exhibit strong co-movement with the market, with betas consistently exceeding 1. Remarkably, eleven of the fifteen analyzed assets outperformed the S&P 500 benchmark. Given that the sample was intentionally restricted to firms and ETFs with a pronounced AI orientation, this widespread outperformance suggests that exposure to the AI theme coincided with exceptional returns over the observation period.

While the total growth rates are noteworthy, Table 20 provides a more granular view of the mean and standard deviation of daily log returns. Here average daily returns cluster around approximately -0.1% to $+0.2\%$.[138] Although these values appear modest in isolation, as indicated by total growth, they compound substantially over time. At the same time, the standard deviations are an order of magnitude larger, indicating significant dispersion and short-term volatility.

Next, three of the four underperforming assets are ETFs, which is noteworthy given that, in theory, diversified portfolios should on average yield more stable and potentially superior risk-adjusted returns than individual stocks (see Section **??**).

This deviation underlines that the selected assets—particularly the individual stocks—represent extreme observations. Hence, while the results are informative for studying narrative-driven dynamics in this specific period, they should be interpreted with caution and not generalized to the broader equity market or other periods.

Elevated β coefficients for NVIDIA (2.18), Tesla (2.33), AMD (2.05), and Broadcom (2.00) indicate a strong sensitivity to market-wide fluctuations. Yet, their compounded α values suggest contrasting outcomes: despite similar systematic risk exposure, AMD and Tesla generated negative abnormal returns, whereas NVIDIA and Broadcom exhibited pronounced positive deviations from CAPM-predicted performance.

AMD—a direct competitor of NVIDIA in semiconductor design—has so far been perceived as one of the relative "losers" in the ongoing "AI race"" (Holzki, 2025).[139] Tesla, in turn, faced a series of challenges, including negative investor sentiment following Elon Musk's increased political engagement and associated reputational controversies (Orr & Inagaki, 2025; Smith & Bushey, 2025).

The ETFs —notably ARKQ, AIQ, and IRBO—also exhibit elevated betas between 1.3 and 1.6. Their relatively high R^2 values (strictly above 70%) indicate that most of their variation is explained by general market dynamics rather than idiosyncratic factors.

Meta, Broadcom, Alphabet, NVIDIA, and TSMC all exhibit relatively low explanatory power ($R^2 \leq 50\%$) despite total growth rates above 100%. This pattern suggests that a large portion of their returns stemmed from idiosyncratic, possibly narrative- or innovation-driven factors rather than systematic market co-movement. It will therefore be particularly interesting to test whether these stocks exhibit stronger linkages to the constructed NLP series in the subsequent analysis.

Apple's under-performance relative to the market is striking in this context, as the firm has been widely portrayed as lagging competitors in integrating artificial intelligence into

[138] Remember that log returns, as long as being close to zero, are roughly one hundredth of percentage change, see 115.

[139] Interestingly, at the time of writing, OpenAI had announced significant investments in AMD, which boosted the company's stock price significantly (https://www.finanzen.net/chart/amd).

its product ecosystem (**leswigAppleFacingPressure2025**; **gallhagerApplesAIEvolution2024**).[140]

Finally, Table 16 in Appendix B summarizes the average daily log returns (in %). The figures again highlight Apple's underperformance: while its mean daily return for 2023–2024 (0.097 %) roughly matches the sample median, it records a notable daily loss of –0.20 % in 2025, yielding only ≈ 0.01 % for 2024–2025 on average. In contrast, NVIDIA unsurprisingly stands out with the highest returns—around 0.30 % per day across the full sample. The dispersion across both time and assets is substantial: vertically, the first three rows reveal marked shifts in performance between years. Horizontally, the range of average returns underscores how unevenly gains were distributed across firms and ETFs. Overall, the table illustrates a mixed and highly heterogeneous sample.

6.3 Granger causality analysis

The central finding is the pervasive absence of statistical significance across most specifications. At the 10% significance level (90% confidence threshold) and without controls, only 103 of 2,790 tests (3.69%) reach significance after Benjamini–Hochberg adjustment of analytical and bootstrapped p-values. This pattern remains unchanged when accounting for semiconductor-sector dynamics via inclusion of the PHLX Semiconductor Index (SOX) (As mentioned above, control variables and asset prices entered the regressions in logarithmic first differences. For ease of exposition, these transformed series are referred to by their respective variable names.). Including the S&P 500 as a control for market-wide movements yields 99 (3.55%) significant tests - a negligible difference.

Including lagged VIX in the model raises the share of AINI→return rejections to 111 (3.98%). Although this increase appears small in relative terms, it corresponds to roughly a one-third rise in absolute rejections for the custom model compared to the S&P500-controlled as a baseline (see Table 5). This warrants closer examination.

Three mechanisms might explain this improvement: (i) a reduction in residual noise due to controlling for a confounder (reduction in Omitted Variable Bias (OVB)); (ii) a decrease in collinearity, improving model conditioning; or (iii) an artificial inflation of significance due to collider bias. The next section tests each of these possibilities successively.

First, OVB: if the VIX affects logged returns and co-moves with AINI, omitting it biases the estimated AINI lag coefficients $\boldsymbol{\gamma}$. According to the canonical OVB formula, $\text{Bias}(\hat{\beta}_{AINI}) = \beta_{\text{VIX}} \frac{\text{Cov}(AINI,\text{VIX})}{\text{Var}(AINI)} > 0$ (Wooldridge, 2010, p. 61 ff.), implying that omitting VIX *overstates* the AINI effect. Including VIX thus purges this confounding channel, reduces the residual variance σ^2, and typically raises the adjusted R^2 by improving model fit.

However, this effect is likely very small: While the VIX exhibits influence on ($\zeta_{\text{VIX}} \neq 0$, see ζ in table **??**), the covariance with the AINI ($\text{Cov}(AINI, \text{VIX})$) is very close to 0 (see table **??**).

Second, improved conditioning: controlling for VIX removes shared variation of lagged AINI and log returns, which can reduce standard errors and increase the precision of $\boldsymbol{\gamma}$. Formally, both σ^2 and the structure of the design matrix determine the variance of

[140] This observation does not imply causal evidence that Apple's stock performance was directly determined by its position in the AI sector. It just indicates that the performance aligns with common interpretations.

the estimator γ and the power of the joint Wald test.[141] While adding a new regressor does not change raw pairwise correlations, it can lower relevant pairwise correlations of regressors and thus improve numerical conditioning (Dolado & Lütkepohl, 1996).[142]

These two mechanisms increase the non–centrality parameter of the joint Wald test for AINI lags, thereby increasing statistical power and thus the rejection rate.

However, as briefly noted in subsection 5.3, a potentially problematic channel remains-namely, the case where the VIX acts as a *collider* rather than a confounder.

In this case, controlling for VIX can induce a spurious correlation between AINI and returns by opening a backdoor path. Intuitively, this occurs when AINI itself increases market-wide uncertainty (raising the VIX), while returns simultaneously affect the VIX, so that movements in the index reflect joint reactions rather than a causal link between AINI and log returns (Elwert & Winship, 2014). [143]

To rule out this possibility, the same Granger-causality framework as above was re-estimated with the VIX as the dependent variable, thereby testing whether AINI Granger-causes the VIX (see equation 22). Table **??** in the appendix reports results for all regression pairs that were significant only in the VIX-controlled specification, i.e., those potentially affected by collider bias. As shown in the last column, none of these regressions are significant, suggesting that collider bias can be safely ruled out.

For robustness, the analysis was repeated over the full time horizon, including all periods, AINI variants, and lag dimensions (a total of 2,700 tests). Again, no significant results emerged (see Table X in the online appendix).

Given the evidence, it looks like including the VIX mainly reduced col-linearity and - albeit to a small extent - reduced OVB. While again the slight increase looks trivial, it will be interesting for certain periods (see below).

We therefore proceed using regressions that include the VIX as a control variable, as its inclusion appears to increase statistical power rather than introduce a spurious backdoor path.

---

[141] The design matrix X collects all regressors - in this case, lagged values of AINI, logged returns, and log growth rates of the VIX.

[142] In linear regression with one omitted confounder Z, the OVB is given by $\text{Bias}(\hat{\gamma}_X) = \zeta_Z \frac{\text{Cov}(X,Z)}{\text{Var}(X)}$, where $X = AINI^{\text{var}}$ and $Z = \text{VIX}$. In matrix notation, $\text{Bias}(\boldsymbol{\gamma}_X) = (X'X)^{-1}X'Z\,\boldsymbol{\zeta}_Z$ holds element-wise (Wooldridge, 2010, p. 61 ff.). For the joint Wald test $H_0 : R\beta = 0$, the test statistic under the alternative is $\lambda = (R\beta)^\top\big[R(X'X)^{-1}R^\top\big]^{-1}(R\beta)/\sigma^2$, with $\text{Var}(\hat{\beta}_j) = \sigma^2[(X'X)^{-1}]_{jj}$. High $\text{Corr}(AINI_{t-j}, r_{t-i})$ or $\text{Corr}(AINI_{t-j}, \text{VIX}_{t-k})$ inflates $[(X'X)^{-1}]_{jj}$, reducing power. Adding VIX lags can reduce partial collinearity, lowering σ^2 and improving the condition number of X. (The condition number is the ratio of largest to smallest eigenvalue of the squared matrix $(X'X)$. A large condition number results in an almost-singular matrix, making it numerically unstable and statistically unreliable (Belsley et al., 2005, p. 100 ff.)). Both effects increase the non-centrality parameter λ, shifting the alternative F-statistic F_{H1} rightward relative to its null distribution and thus increasing rejection probability (Dolado & Lütkepohl, 1996). This reasoning follows the classical OLS framework where the Wald statistic's non-centrality parameter λ determines power. Note that in this empirical setting, which uses Newey–West variance estimators and bootstrap-based inference, while the exact distribution is non-standard, this still holds from a geometrical perspective (Kiefer & Vogelsang, 2005; MacKinnon, 2009). Thus, the argument should be viewed as heuristic rather than formally exact.

[143] Formally, under linearity and joint normality, $\text{Cov}(AINI_t, r_t \mid VIX_t) = -\frac{\text{Cov}(AINI_t, VIX_t)\,\text{Cov}(r_t, VIX_t)}{\text{Var}(VIX_t)}$. Hence, even if $\text{Cov}(AINI_t, r_t) = 0$, conditioning on VIX can induce nonzero dependence through their common effect on the index (Elwert & Winship, 2014; Pearl, 2009).

Figure 6 visualizes the number of significant Granger causality results (See Appendix II, Table 11 for full table, with total number of possible rejections given in parenthesis) by ticker and period. Each bar represents the number of jointly rejected null hypotheses ($\alpha = 0.1$).

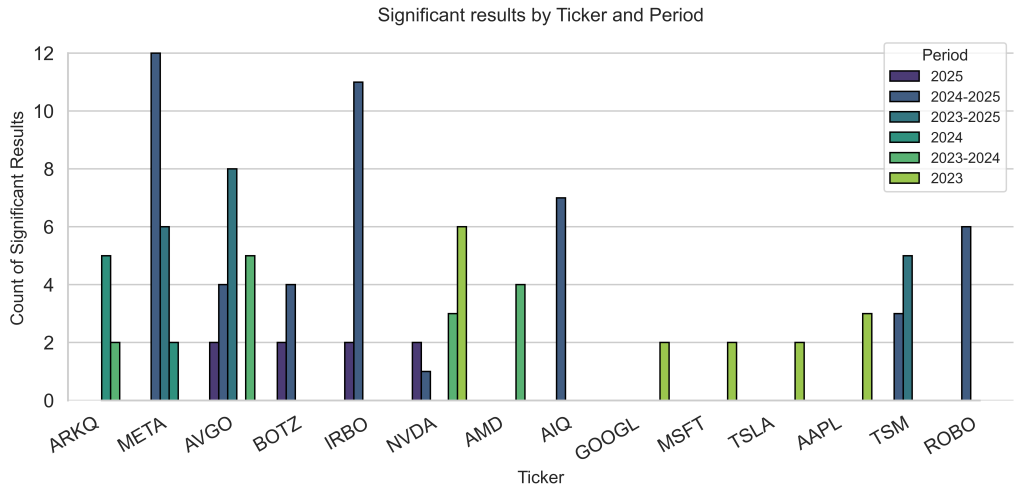

Figure 6. Number of jointly rejected Granger causality tests (controlled for VIX) per ticker and period at $\alpha = 0.1$. *Source:* Own illustration.

Across ticker–period pairs, up to 36 Granger tests were estimated (three measures × four model variants × three lags). In 2025, the count dropped to nine due to non-stationarity. Even the strongest case - Meta (2024–2025) - yielded only 12 of 36 rejections (33%), comparable in relative terms to Broadcom, BOTZ, NVIDIA, and IRBO (each 2 of 6, 33%) in 2025.

Ranking all assets by total rejections yields the results in Table 4, with total number of possible rejections given in parenthesis. Interestingly, four of the five most-responsive assets (META, Broadcom, NVDIA, TSMC) intersect with those of the highest abnormal returns Table 19). From top-performers, Amazon is the exception, displaying zero significant results despite strong returns. Notably, the iShares Future AI & Tech ETF (IRBO), while under-performing the market, exhibits *relatively* high sensitivity to the narrative indices.

Ticker	Jointly Rejected at $\alpha = 0.1$
META	20 (186)
AVGO	19 (186)
IRBO	13 (186)
NVDA	12 (186)
TSM	8 (186)
AIQ	7 (186)
ARKQ	7 (186)
BOTZ	6 (186)
ROBO	6 (186)
AMD	4 (186)
AAPL	3 (186)
GOOGL	2 (186)
MSFT	2 (186)
TSLA	2 (186)

Table 4. Number of jointly rejected Granger causality tests per Ticker (AINI → Return, controlled for VIX) at $\alpha = 0.1$ across all periods. *Source:* Own calculations. Table prepared with the assistance of generative AI (https://chatgpt.com/share/68df816e-d7b4-8013-874b-ccd146b5823b).

Finally, Table 5 summarizes rejections by model variant, with total counts from the S&P500-controlled setup reported in parentheses and total number of tests ran in square brackets. The overall pattern is straightforward: naïve, dictionary-based models with narrower context windows yielded more rejections than the custom variant. Given the high agreement in AI-relatedness between the naïve and custom classifications (97.9%, see Section 5.1), the weaker relative performance of the custom model likely reflects its substantially broader context window, which diluted signal relevance by capturing sentiment not directly related to *AI* (see Example 1 in Section 6.1). This under-performance is particularly noteworthy since the model-applied to an additional 90 regression pairs, as indicated by the numbers in square brackets-was tested more extensively than the other variants.

Among the window-based variants, rejection rates increased monotonically with window size. Interestingly, all models except w_1 improved when including the VIX, suggesting that most AINI variants implicitly captured uncertainty-related dynamics also reflected in the volatility index.

Model	Number of Joint Rejections, incl. S&P500
w_2	56 (51) [675]
w_1	21 (21) [675]
w_0	18 (15) [675]
custom	16 (12) [765]

Table 5. Number of jointly rejected Granger causality tests per Ticker (AINI → Return, controlled for VIX (S&P500)) at $\alpha = 0.1$. Number of tests run in square brackets. *Source:* Own calculations.

Thus, it appears that excessively broad textual contexts introduce semantic noise that weakens the predictive signal, while overly narrow windows may fail to capture sufficient contextual information.[144]

While table 5 may suggest that w_0 is largely a subset of w_1 and both of w_2, it is useful to formally assess how these sets of significant observations align. For this purpose, the *Jaccard index*, denoted by J, provides a useful measure of pairwise similarity (Costa, 2021). It quantifies the similarity between two sets A and B as the ratio of their intersection over union:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{25}$$

In this context, each set A and B corresponds to the distinct combinations of ($AINI^{variant}$, Ticker, Period, Lag) that were jointly significant for two models simultaneously.[145]

The matrix below reports those pairwise Jaccard indices. Values on the diagonal equal 1 by construction, since each model is perfectly similar to itself, while off-diagonal entries capture the overlap between different models.

	custom	**w0**	**w1**	**w2**
custom	1.000	0.179	0.034	0.054
w0	0.179	1.000	0.125	0.098
w1	0.034	0.125	1.000	0.178
w2	0.054	0.098	0.178	1.000

Table 6. Jaccard Index across all dimensions. *Source:* Own calculations.

[144] Determining the optimal context window lies beyond the scope of this paper. Preliminary experiments indicated a sharp decline in significance when regressions were estimated without the bootstrap procedure described above. As a single full pass through the inference pipeline-including sentiment estimation and all robustness checks-requires roughly four to five days on the available hardware, a more exhaustive investigation was not feasible within the current project.

[145] Simply speaking, an overlap is only accounted as an overlap if two models reject H_0 for the same period, ticker and lag combination.

With $J(w_0, custom) = 0.179$ being the largest overlap, it becomes evident that each model seems to have captured narratives and sentiment for a different subset of stock-period-lag pairs.

Examining this issue further, Table 7 investigates overlaps in estimated γ_1 coefficients across identical Ticker-Period-Lag combinations No Significant Results (NSR) denotes cases without significant Granger causality across all models. The first column lists the unique Ticker-Period-Lag identifiers, with the maximum lag length indicated in parentheses. Subsequent columns display β_1 estimates for each model variant as labeled.

Ticker / Period / Lags	custom	w0	w1	w2
AAPL 2023 (2)	NSR	NSR	NSR	-0.012433
AAPL 2023 (3)	NSR	NSR	NSR	-0.015545
AIQ 2024–2025 (2)	NSR	NSR	-0.024342	-0.025747
AIQ 2024–2025 (3)	NSR	NSR	NSR	-0.024790
AMD 2023–2024 (1)	NSR	NSR	NSR	-0.019922
AMD 2023–2024 (3)	NSR	-0.010485	NSR	NSR
ARKQ 2023–2024 (3)	NSR	-0.010934	NSR	NSR
ARKQ 2024 (1)	-0.011315	NSR	NSR	NSR
ARKQ 2024 (2)	-0.010104	NSR	NSR	NSR
AVGO 2023–2024 (1)	-0.014943	NSR	NSR	NSR
AVGO 2023–2024 (3)	NSR	-0.005818	NSR	NSR
AVGO 2023–2025 (1)	0.016757	NSR	NSR	NSR
AVGO 2023–2025 (2)	NSR	NSR	NSR	0.013875
AVGO 2023–2025 (3)	NSR	0.021368	NSR	0.015790
AVGO 2024–2025 (2)	NSR	NSR	NSR	0.020637
AVGO 2024–2025 (3)	NSR	NSR	NSR	0.019833
AVGO 2025 (1)	0.100553	NSR	NSR	NSR
BOTZ 2024–2025 (2)	NSR	NSR	NSR	-0.014038
BOTZ 2024–2025 (3)	NSR	NSR	NSR	-0.013645
BOTZ 2025 (1)	0.008359	NSR	NSR	NSR
GOOGL 2023 (1)	NSR	NSR	NSR	-0.005268
IRBO 2024–2025 (2)	NSR	NSR	-0.009017	-0.008642
IRBO 2024–2025 (3)	NSR	NSR	-0.007561	-0.009228
IRBO 2025 (1)	0.042440	NSR	NSR	NSR
META 2023–2025 (1)	NSR	NSR	0.002531	0.002068
META 2023–2025 (2)	NSR	NSR	0.003022	NSR
META 2024 (1)	-0.020182	NSR	NSR	NSR
META 2024 (2)	-0.018587	NSR	NSR	NSR
META 2024–2025 (1)	0.003266	0.004136	0.003484	0.002946
META 2024–2025 (2)	NSR	0.005273	0.003870	0.003291
MSFT 2023 (1)	NSR	NSR	NSR	0.003800
NVDA 2023 (1)	NSR	NSR	NSR	0.006610
NVDA 2023 (2)	NSR	NSR	NSR	0.013224
NVDA 2023 (3)	NSR	NSR	NSR	0.014381
NVDA 2023–2024 (3)	NSR	-0.004490	NSR	NSR
NVDA 2024–2025 (1)	-0.004193	NSR	NSR	NSR
NVDA 2025 (1)	0.058003	NSR	NSR	NSR
ROBO 2024–2025 (2)	NSR	NSR	NSR	-0.012454
ROBO 2024–2025 (3)	NSR	NSR	NSR	-0.012602
TSLA 2023 (1)	NSR	NSR	NSR	-0.097526
TSM 2023–2025 (2)	NSR	NSR	NSR	-0.003251
TSM 2023–2025 (3)	NSR	NSR	NSR	-0.002992
TSM 2024–2025 (2)	NSR	NSR	NSR	-0.014394
TSM 2024–2025 (3)	NSR	NSR	NSR	-0.016952

Table 7. Overlap in γ_1 coefficient estimates across Ticker–Period–Lag combinations. *Source:* Own calculations. Prepared with the assistance of generative AI (https://chatgpt.com/share/68f4c521-9f98-8013-b29e-2523c873bf1b).

Reading the table from left to right, only META (2024–2025) exhibits significance across all model variants simultaneously, reaffirming that each specification captures different signals. This holds true even when extending the analysis across all periods, i.e. when taking into account to which models a given stocks exhibited sensitivity on all periods.

Consistent with the overall rejection counts by ticker discussed above, TSMC, ROBO,

Tesla, Microsoft, and Apple exhibit significance exclusively under w_2, while no other model shows such dominance. By contrast, NVIDIA (5/6), BOTZ (3/3), and Broadcom (7/9) display a clear bias toward models with broader contextual scope (w_2 and *custom*). Several tickers – including Apple, AIQ, AMD, ARKQ, BOTZ, ROBO, Tesla, and TSMC – are characterized predominantly by negative coefficients, whereas others show periodic positive relationships with the AINI, indicating heterogeneous temporal sensitivities.

These patterns do not align neatly with realized growth rates. META, TSMC, NVIDIA, and Broadcom – all among the best-performing stocks in terms of cumulative returns – exhibit distinct sensitivities across models, time horizons, and even coefficient signs. Equally notable is the mostly absence of significance for Microsoft, despite its central role in the recent AI boom and its strategic partnership with OpenAI (Hagey, 2025). Given that OpenAI was explicitly included as a filter term in the article selection process (see Section 5.1), this lack of sensitivity is particularly striking. We will return to this issue in the discussion section.

It is also instructive to examine the distribution of regression coefficients across models and different lag length. Figure 7 displays histograms of γ_1, γ_2, and γ_3 for each model, with the corresponding means denoted by $\bar{\gamma}$.

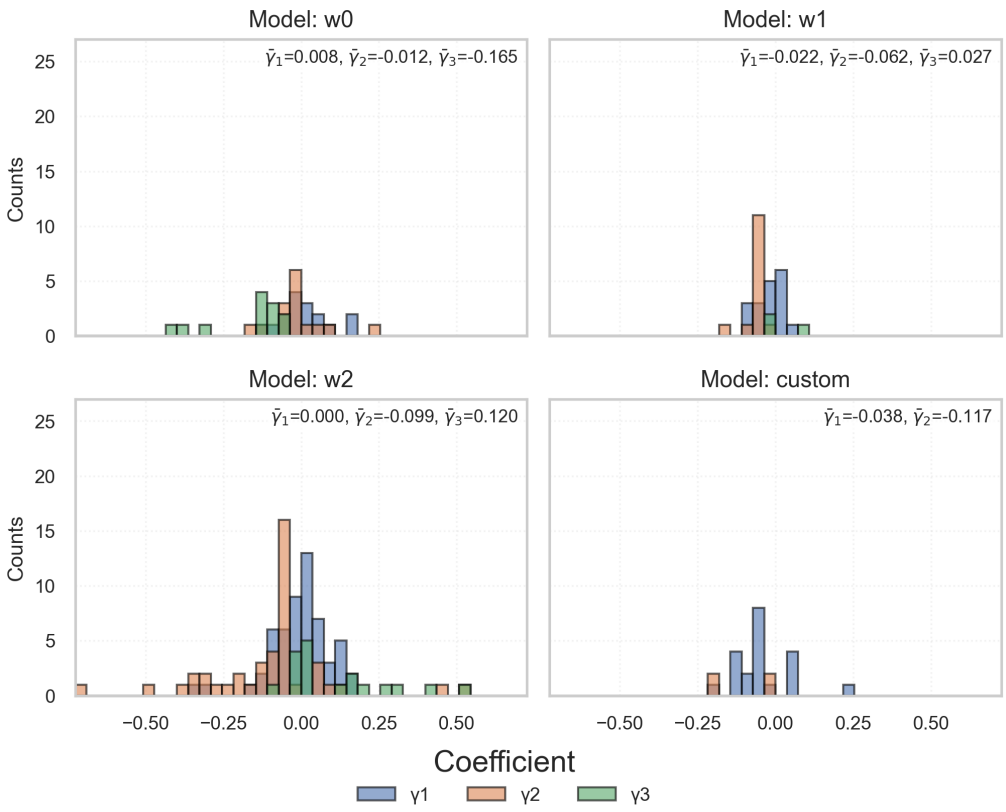

Figure 7. Distribution of γ coefficients across models and lags. *Source:* Own illustration.

The distribution reveals distinct patterns across models. For γ_1 and γ_3, means change signs across models, while $\bar{\gamma_2}$ is always slightly negative. However, in general, all regression coefficients a centreded towards 0 - indicating the general minor magnitudes of even significant results.[146]

[146] As expected under finite variance and approximate normality of estimators - facilitated by using log-

Sorting the jointly significant results by absolute magnitude of γ_1 (full table in Appendix, Table 12; all 3,735 regressions in the online appendix `gc_VIX_aini_to_ret.html`) yields the following:

Granger–Causality: jointly significant results (AINI → Returns, VIX-controlled)											
Model	$AINI^{\text{var}}$	Period	Ticker	β_1	ζ_1	γ_1	γ_2	γ_3	R^2_{adj}	p_a^{BH}	p_e^{BH}
w2	$EMA^{0.2}$	2023	NVDA	0.027	0.014	-0.334	0.535	-0.074	0.031	0.02**	0.02**
w2	$EMA^{0.2}$	2023	NVDA	0.021	0.013	-0.312	0.459		0.032	0.01**	0.01***
custom	$EMA^{0.2}$	2024	ARKQ	-0.060	-0.010	0.242	-0.212		0.023	0.03**	0.05**
custom	$EMA^{0.2}$	2024	META	-0.111	-0.020	-0.191			0.031	0.04**	0.07*
w2	$EMA^{0.2}$	2024-2025	AVGO	0.017	0.020	0.181	-0.726	0.511	0.024	<0.01***	0.01***
w0	$EMA^{0.2}$	2023-2024	AVGO	0.054	-0.006	0.158	-0.064	-0.321	0.014	0.01***	0.01***
w2	$EMA^{0.2}$	2024-2025	BOTZ	-0.123	-0.014	0.156	-0.222		0.015	0.04**	0.06*
w0	$EMA^{0.2}$	2023-2025	AVGO	0.018	0.021	0.148	-0.166	-0.121	0.008	0.06*	0.08*
w2	$EMA^{0.2}$	2023-2024	AMD	-0.039	-0.020	-0.146			0.001	0.05*	0.07*
w2	$EMA^{0.2}$	2024-2025	BOTZ	-0.108	-0.014	0.143	-0.337	0.143	0.021	0.08*	0.03**
w2	$EMA^{0.2}$	2024-2025	AIQ	-0.203	-0.026	0.141	-0.200		0.023	0.03**	0.06*
custom	$EMA^{0.2}$	2024-2025	NVDA	-0.137	-0.004	-0.140			0.018	0.05*	0.08*
w2	$EMA^{0.2}$	2024-2025	ROBO	-0.107	-0.012	0.139	-0.194		0.014	0.03**	0.06*
w0	$EMA^{0.2}$	2024-2025	META	-0.086	0.004	-0.134			0.011	0.04**	0.03**
w2	$EMA^{0.2}$	2024-2025	ROBO	-0.092	-0.013	0.133	-0.323	0.151	0.021	0.04**	0.03**
w2	$EMA^{0.2}$	2024-2025	AIQ	-0.183	-0.025	0.130	-0.319	0.145	0.031	0.06*	0.03**
custom	$EMA^{0.8}$	2025	NVDA	-0.128	0.058	-0.124			0.053	0.01***	0.05**
w2	$EMA^{0.8}$	2023	TSLA	0.005	-0.098	-0.122			0.031	0.03**	0.08*
custom	$AINI^{norm}$	2025	NVDA	-0.131	0.059	-0.114			0.054	0.01***	0.05**
custom	$EMA^{0.8}$	2025	AVGO	-0.075	0.101	-0.111			0.081	0.08*	0.08*

Table 8. Granger–Causality, jointly significant results (AINI → Returns, controlled for VIX). Top 20 sorted by $|\gamma_1|$. *Source:* Own. Signif.: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

When ordering the jointly significant regressions by the absolute magnitude of the first-lag coefficient on the AINI ($|\gamma_1|$), a distinct pattern emerges. Sixteen of the twenty strongest effects stem from specifications using $EMA^{0.2}$, while only one regression (second-to-last entry) relies on a non-EMA variant of the index. This concentration suggests that the exponentially weighted specification with a 0.2 decay factor either better captured the persistence structure of AI-related sentiment or more effectively smoothed out calendar-induced noise associated with weekends and banking holidays.

While models with shorter text windows (w_0, w_1) perform relatively well in terms of the overall number of rejections (see Table **??**) compared with the *custom* specification, the latter is disproportionately represented among the largest coefficient magnitudes, whereas w_1 is entirely absent from the top twenty. Apart from a few exceptions, the leading results are generally dominated by large context windows (w_2 and *custom*), suggesting that richer contextual information enhances the informational content of the sentiment measure.

Across stocks, the pattern broadly mirrors earlier findings. NVIDIA, META, and Broadcom exhibit the highest sensitivity to AI-related sentiment both in terms of coefficient magnitude and frequency of significance. The AI-focused ETFs also appear prominently, implying that sector-level indices systematically co-move with AI narratives, despite their underperformance relative to the broader market benchmark.

When both γ_1 and γ_2 are estimated, their signs are typically opposite, indicating pronounced non-monotonicity in the dynamic response. Among all 101 significant regres-

transformed, stationary series - the empirical distribution of the estimated γ coefficients in the model with the largest number of 'draws' (w_2) already roughly resembles the canonical t-distribution (Wooldridge, 2010, p.53 ff.). This aligns with the asymptotic theory behind the Wald test, which for single restrictions is equivalent to squared t-statistics, and whose geometric intuition extends to multiple restrictions in higher dimensions, in both its analytical and bootstrapped implementations. However, this visual pattern should not be overstated, as the plotted coefficients represent only the subset of significant results, which is not expected to follow the theoretical sampling distribution (see footnote 142) and Kiefer and Vogelsang (2005) and MacKinnon (2009).

sions, 76 include more than one lag of AINI, and within these, 57 (75%) exhibit a sign switch between γ_1 and γ_2. Regressions with same-sign coefficients tend to cluster in the lower quartile of $|\gamma_1|$, suggesting that mean-reverting dynamics dominate the stronger effects. Single-lag models still account for 8 of the 20 largest $|\gamma_1|$ estimates, possibly because they absorb lagged components of the sentiment shock. For instance, META (2024, $EMA^{0.2}$) remains significant under a three-lag specification, where all AINI coefficients are negative, consistent with short-lived reactions followed by partial reversal.

In terms of magnitude, the dispersion of γ_1 is substantial, ranging from -0.334 (NVIDIA 2023, $EMA^{0.2}$) to approximately zero. The autoregressive return coefficient (β_1) is generally smaller in absolute value and frequently below γ_1, while the contemporaneous VIX control (ζ_1) is typically weaker still. The distribution of R^2_{adj} is highly right-skewed: most regressions explain only about 1% or less of daily return variation, yet a small subset achieves considerably higher explanatory power-up to 8% for Broadcom (2025, *custom* $EMA^{0.8}$ or normalized AINI) and around 5% for META (2025, *custom* $EMA^{0.2}$). The five highest R^2_{adj} values all arise from the *custom* window, with w_1 first appearing around rank 25 and w_0 only around rank 36.

To build intuition, it is useful to recall what the coefficient γ_1 represents. It measures the expected change in the daily log return r_t following a one-unit change in the given AINI. Since log returns were defined as first differences of adjusted log prices, a value of $r_t = 0.01$ corresponds roughly to a 1% daily price increase (see footnote 115). A one-unit increase in r_t would thus correspond to a 100% price jump.

Given mean AINI values of -0.005 for w_2 and 0.09 for the custom variant (see Table 1), a coefficient of $\gamma_1 = -0.1$ - which lies below the top-20 magnitudes - implies an expected daily change in returns of $-0.1 \times 0.005 = -0.0005$, i.e., about -0.05% for w_2 and $-0.1 \times 0.009 = -0.0009$ ($\approx -0.09\%$) for the custom model. This effect size exceeds roughly 5 to as much as 35 times the standard deviation of daily log returns across the sample assets (see Table 20).[147] Hence, while γ_1 coefficients may appear modest numerically, their implied short-term return responses are economically meaningful.

Finally, a large fraction of coefficients are significant only at the 10% level, reflecting the high volatility and noise inherent in daily data. Notably, in 29 of the 101 regression pairs, the analytical p-value falls below its bootstrapped counterpart, implying that in roughly one third of cases, the analytical test acted as the more conservative benchmark, while the empirical (bootstrapped) did so in the remaining cases - underscoring the importance of employing both inference procedures.

It is also useful to briefly contrast these results with the opposite direction, i.e., when returns act as the independent variable as defined in equation 21. As above mentioned, 111 out of 2,790 tests (3.98%) were significant for the channel *AINI* $\rightarrow$ *Returns*, with the same baseline, the number of rejections increases to 146 ($\approx 5.23\%$) when reversing the direction. The overlap between both remains limited: Ignoring lag-dimensions, AINI variants and models, i.e. focusing on identical combinations of periods and tickers, there is a set of 23 unique AINI $\rightarrow$ Returns combinations and 26 in the reversed direction, but a total overlap of only 3, i.e. there are only three period-ticker pairs significant for both.[148]

[147] Note that this is a very rough intuition, based on conservative rounding.

[148] Microsoft in 2023, Apple in 2023, and META in 2024.

As seen below, in this case the regression-coefficients are not the most-striking fact, but rather R^2_{adj} for some regression pairs. Thus, the table below is sorted by those, representing again the top 20 results.

Granger–Causality: jointly significant results (Returns → AINI, VIX-controlled)										
Model	$AINI^{\text{var}}$	Period	Ticker	β_1	β_2	β_3	γ_1	R^2_{adj}	p_a^{BH}	p_e^{BH}
w1	$EMA^{0.2}$	2024_25	ROBO	0.132			0.928	0.847	0.08*	0.09*
w2	$EMA^{0.2}$	2024_25	ROBO	0.147	-0.087		0.875	0.828	0.05*	0.04**
w2	$EMA^{0.2}$	2024_25	BOTZ	0.105	-0.101		0.875	0.828	0.04**	0.06*
w2	$EMA^{0.2}$	2024_25	AMD	0.011	-0.054		0.868	0.828	0.07*	0.03**
w2	$EMA^{0.2}$	2024_25	AMZN	0.007	-0.089		0.865	0.827	0.04**	0.09*
w2	$EMA^{0.2}$	2024_25	BOTZ	0.103	-0.100	0.038	0.876	0.826	0.08*	0.10*
w2	$EMA^{0.2}$	2024_25	ROBO	0.145	-0.087	0.018	0.873	0.826	0.09*	0.07*
w2	$EMA^{0.2}$	2024_25	ROBO	0.155			0.918	0.824	0.04**	0.04**
w1	$EMA^{0.2}$	2023_24_25	META	-0.054	-0.048		0.889	0.791	0.03**	0.03**
w1	$EMA^{0.2}$	2023_24_25	META	-0.053	-0.048	-0.019	0.888	0.791	0.05*	0.06*
w1	$EMA^{0.2}$	2023_24_25	META	-0.047			0.888	0.789	0.04**	0.09*
custom	$EMA^{0.2}$	2024_25	BOTZ	0.133			0.871	0.755	0.02**	0.08*
w2	$EMA^{0.2}$	2023_24_25	AMZN	-0.008	-0.064		0.845	0.754	0.06*	0.09*
w2	$EMA^{0.2}$	2023_24_25	META	-0.024	-0.046		0.840	0.754	0.05*	0.10*
w2	$EMA^{0.2}$	2023_24_25	META	-0.025	-0.048	-0.032	0.832	0.752	0.02**	0.07*
w1	$EMA^{0.2}$	2024	META	-0.065			0.839	0.713	0.04**	0.07*
custom	$EMA^{0.2}$	2023_24_25	BOTZ	0.166	-0.067	0.118	0.964	0.713	0.03**	0.02**
custom	$EMA^{0.2}$	2023_24_25	BOTZ	0.183	-0.064		0.954	0.711	0.03**	0.03**
custom	$EMA^{0.2}$	2023_24_25	NVDA	0.043	0.004	0.055	0.958	0.711	0.04**	0.06*
w1	$EMA^{0.2}$	2024	META	-0.067	-0.022		0.782	0.711	0.10*	0.08*

Table 9. Granger–causality, jointly significant results (Returns → AINI, controlled for VIX). Sorted by R^2_{adj}. *Source:* Own calculations. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In total, 49 out of the 146 estimated regressions exhibit an adjusted R^2 above 50%, with the highest value approaching 85%. The w_2 specification again dominates among the top-performing models, although the relative prominence of the custom variant declines somewhat. Notably, the single best-fitting regression is obtained under the w_1 configuration. Interestingly, several ETFs - particularly ROBO and BOTZ - account for a large share of the models with high explanatory power, despite their already-discussed weaker market performance.

Next, the estimated coefficients on lagged returns (β) appear non-trivial, with several values in the range of 0.5–0.95. As above, it is crucial to interpret these magnitudes in light of scale. In the data, even the largest observed daily returns are below $|r_t| \approx 0.004$ (0.4%), as reported in Table 20. Taking the upper-end estimate $\beta_1 = 0.95$ and combining it with the largest mean daily return of 0.004 implies an expected daily change in AINI of only $0.95 \times 0.004 \approx 0.0038$. This equals roughly 8% of the index's standard deviation ($\sigma_{AINI} \approx 0.045$, see Table 1) for the most-extreme combination. Hence, despite appearing numerically large, the β coefficients correspond to small effects in absolute terms. Regarding sign-consistency, of 109 regression-pairs with at least 2 lags, 50 ($\approx 46\%$) show a sign-switch.

Finally, the smoothing pattern roughly mirrors that observed on the AINI→Returns side *when sorting by explained variance*. Interestingly, when ranking by the absolute magnitude of β_1 (see Table 13 in the appendix), the memoryless $AINI^{norm}$ and the short-term memory specification $EMA^{0.8}$ emerge as the dominant variants.

After presenting the descriptive evidence and regression outcomes, the following section interprets these findings through the lenses of behavioral finance and classical no-arbitrage theories.

7 Discussion

Revisiting the initial setup, two hypothesis were formulated: (H1) narratives Granger-cause short-horizon price changes of the same sign, (H2) these effects reverse at later lags (mean reversion).

In general, both H1 and H2 must be rejected. The vast majority of Granger-causality tests yielded insignificant results, suggesting that the AINI did not meaningfully Granger-cause (log) returns during the examined periods. From a scientific standpoint, the central question is therefore *why*.

A first explanation lies in what might be termed an *NLP version* of the joint-hypothesis problem.[149] As A. H. Huang et al. (2023, p. 824) note, any test of return predictability based on sentiment is simultaneously a test of whether the NLP model accurately captures textual sentiment *and* whether the market reacts to such sentiment in the expected way.

It is therefore useful to distinguish between two critical components of the sentiment inference process: (i) identifying whether an article contains narratives related to AI, and (ii) estimating the sentiment of that article.

Given the extensive evidence that BERT and its domain-specific derivative FinBERT perform strongly on benchmark datasets and unseen data (Acheampong et al., 2021; Devlin et al., 2019; Jiang & Zeng, 2023; Yang et al., 2020), it is reasonable to assume that component (ii) is not the primary source of error. The key challenge is therefore whether AI-related narratives were identified reliably in the first place, either through the naïve keyword dictionary or via the fine-tuned FinBERT classifier.

There is some evidence supporting the performance of these models. All specifications - particularly the window-based variants - exhibited pronounced spikes during major AI-related events, such as the roll-out of the EU AI Act (see Figure 5). Both the dictionary- and FinBERT-based methods also performed reasonably well on the manually annotated test data. Moreover, the very high adjusted R^2 values observed for the Returns→AINI channel suggest that the indices successfully captured shifts in narrative attention.[150]

However, as discussed in Section 4, well-established dictionary-based indices such as the EPU or GPR rely on extensive and iterative keyword-selection procedures. It is therefore likely that both the naïve labeling dictionary and the custom FinBERT classifier produced suboptimal results, potentially including a substantial number of false positives, as illustrated in Example 3 of Subsection 6.1. This implies that the resulting AINI remains a relatively noisy estimator.

[149] As discussed in Subsection 2.1, the joint-hypothesis problem in finance refers to the fact that one cannot test market efficiency without simultaneously testing the model of efficiency. A rejection may therefore reflect either true inefficiency or model misspecification.

[150] Regarding the latter, the key question is what might explain this reverse channel. While a comprehensive discussion lies beyond the scope of this thesis, it is plausible that the WSJ's editorial attention is at least partly shaped by stock market dynamics. It is unlikely, however, that the underperforming ETFs themselves directly triggered the WSJ's increased focus on AI. A more reasonable interpretation is that these ETFs proxy for a latent, market-wide interest in AI-related themes, which was subsequently mirrored by the WSJ's coverage—consistent with the broader phenomenon that financial media partly reports on what investors are already discussing (Mullainathan & Shleifer, 2005; Shiller, 2021; Taffler et al., 2024). Taken together, this pattern also lends support to the robustness of the underlying dictionary-based approach.

The true data-generating process behind asset prices-investor behavior is likely another factor. According to the semi-strong form of the EMH, next-day prices should already incorporate any publicly available information (see Subsection 2.1). More general, even without assuming the EMH, as long as being an AI-related company does not constitute a systematic risk factor, the APT also predicts that AINI-based sentiment should not yield persistent excess returns.

Behavioral theory adds an additional layer of complexity: while many studies emphasize noise trading and emotionally driven biases, the *measurement* of such behavior is far from straightforward. As discussed earlier, irrational behavior often arises from vivid experiences and plausible *sounding* narratives rather than from simple frequency and sentiment.

Consequently, the relevant signal may stem from salient, event-driven stories-such as the emergence of DeepSeek or the release of ChatGPT - none of which are well captured by the daily, aggregate AINI.[151]

Furthermore, Shiller (2021) highlights that narratives operate in constellations rather than in isolation, with their influence driven more by interaction than by frequency, and evolving across heterogeneous and often long time scales. Therefore, the absence of short-term predictability does not necessarily imply the absence of narrative influence - it may simply reflect a nonlinear and delayed transmission mechanism between collective beliefs and asset prices.

While the broad-based dictionary approach may capture diverse narrative types-as seen in Examples 1 and 2 of Subsection 6.1, which identified stories about ASML's price surge and Telegram's use of AI to monitor undemocratic behavior-the remaining challenges - in addition to measuring vividness and plausibility- are twofold: (i) how to measure narrative strength when frequency is not the key signal, and (ii) how to account for varying temporal scales of narrative diffusion.

Also, the issue of *data mining* must be acknowledged. Empirical finance faces a vast parameter space: multiple assets, time periods, lag structures, and variable definitions. Even with rigorous methods, it is always possible - by chance or design - to identify subsets of statistical significance (Malkiel, 2003). This problem scales nonlinearly with the growth of computational power and the increasing flexibility of NLP tools. The more sources, tokens, and keyword combinations researchers can test, the greater the risk of finding 'significant' results that do not generalize out of sample.

Returning to the empirical findings, the question becomes whether the observed effects represent mere statistical noise—or whether they capture meaningful yet episodic relationships. The answer is, inevitably, *it depends*.

The preceding analysis provides a detailed assessment of whether narratives surrounding AI in the WSJ contain predictive information for asset prices, while also examining whether market movements shape the narrative environment itself. Across all specifications and subperiods, the evidence points predominantly toward the latter: market

[151] One might argue that distinct peaks in narrative intensity around major events serve as reasonable proxies, yet the question remains how to measure *vividness* precisely. Here, vividness is meant quite literally: following the literature discussed above, the emergence of ChatGPT should have had a large impact, as many individuals directly experienced it in their daily lives, whereas DeepSeek likely remained confined to more specialized circles. Moreover, even temporary surges in media attention do not necessarily capture this lived, day-to-day engagement with AI.

dynamics appear to precede and inform changes in narrative tone and intensity, as indicated by the high adjusted R^2 values in the Returns→AINI direction.

Given the absence of consistent, sign-aligned, and cross-period significance across AINI measures, the conservative conclusion is that most effects reflect noise. However, the occasional significance observed for specific assets or sub-periods suggests that emotionally charged narratives can, at times, exert a real influence on financial markets. For practitioners, this limits the AINI's predictive value.

For researchers, these links invite deeper inquiry: *When and why do certain narratives matter, and how can they be measured more robustly?* Moreover, how can we explain the existence of mean reversion in some but not all cases?

In this sense, the results of this thesis do not falsify the narrative hypothesis but rather illuminate its empirical boundaries. In line with theories of emotional overreaction, the findings show that, in certain periods and for very specific assets, markets react strongly in terms of volatility - when measured by standard deviation - yet not in terms of explained variation. Stated differently, the AINI can Granger-cause transient spikes in returns, but it cannot explain how prices are ultimately determined.

Finally, this study also points toward promising directions for future research. A major obstacle remains the reliable capture of narratives. Future work should therefore focus on leveraging transformer-based architectures - particularly emerging agentic AI models - as annotation tools capable of identifying and classifying narratives more precisely (Gilardi et al., 2023, see). In addition, the analysis could be extended to other media outlets, which likely convey different narrative tones and target different audiences (S. Baker et al., 2016, see). Lastly, non-linearities and interaction effects warrant further attention, potentially through the application of ANN-based frameworks or time-varying parameter models.

List of Equations

List of Figures

List of Tables

A Statistical testing & estimates

Stationarity tests								
Measure	Period	ADF_{stat}	ADF_p	PP_{stat}	PP_p	$KPSS_{stat}$	$KPSS_p$	agree
AAPL	2023	-12.547	0.000***	-12.503	0.000***	0.081	0.100	Yes
AAPL	2023–2024	-19.436	0.000***	-19.386	0.000***	0.093	0.100	Yes
AAPL	2023–2025	-11.906	0.000***	-22.816	0.000***	0.116	0.100	Yes
AAPL	2024	-15.012	0.000***	-15.026	0.000***	0.173	0.100	Yes
AAPL	2024–2025	-10.580	0.000***	-18.913	0.000***	0.127	0.100	Yes
AAPL	2025	-6.435	0.000***	-11.304	0.000***	0.034	0.100	Yes
AIQ	2023	-12.196	0.000***	-12.300	0.000***	0.103	0.100	Yes
AIQ	2023–2024	-19.732	0.000***	-19.719	0.000***	0.035	0.100	Yes
AIQ	2023–2025	-14.104	0.000***	-23.461	0.000***	0.038	0.100	Yes
AIQ	2024	-15.573	0.000***	-15.795	0.000***	0.030	0.100	Yes
AIQ	2024–2025	-19.668	0.000***	-19.776	0.000***	0.036	0.100	Yes
AIQ	2025	-11.219	0.000***	-11.204	0.000***	0.129	0.100	Yes
AMD	2023	-14.422	0.000***	-14.403	0.000***	0.147	0.100	Yes
AMD	2023–2024	-21.325	0.000***	-21.340	0.000***	0.212	0.100	Yes
AMD	2023–2025	-24.525	0.000***	-24.631	0.000***	0.171	0.100	Yes
AMD	2024	-15.934	0.000***	-15.943	0.000***	0.147	0.100	Yes
AMD	2024–2025	-19.973	0.000***	-20.085	0.000***	0.074	0.100	Yes
AMD	2025	-11.657	0.000***	-11.844	0.000***	0.138	0.100	Yes
AMZN	2023	-12.017	0.000***	-15.559	0.000***	0.096	0.100	Yes
AMZN	2023–2024	-17.693	0.000***	-22.020	0.000***	0.063	0.100	Yes
AMZN	2023–2025	-24.581	0.000***	-24.993	0.000***	0.133	0.100	Yes
AMZN	2024	-12.969	0.000***	-15.465	0.000***	0.063	0.100	Yes
AMZN	2024–2025	-19.665	0.000***	-19.837	0.000***	0.079	0.100	Yes
AMZN	2025	-11.823	0.000***	-11.825	0.000***	0.155	0.100	Yes
ARKQ	2023	-11.884	0.000***	-11.886	0.000***	0.117	0.100	Yes
ARKQ	2023–2024	-19.737	0.000***	-19.707	0.000***	0.179	0.100	Yes
ARKQ	2023–2025	-13.176	0.000***	-23.181	0.000***	0.085	0.100	Yes
ARKQ	2024	-15.729	0.000***	-15.755	0.000***	0.367	0.091	Yes
ARKQ	2024–2025	-11.142	0.000***	-19.577	0.000***	0.131	0.100	Yes
ARKQ	2025	-11.022	0.000***	-11.045	0.000***	0.186	0.100	Yes
AVGO	2023	-12.866	0.000***	-12.859	0.000***	0.089	0.100	Yes
AVGO	2023–2024	-8.453	0.000***	-19.531	0.000***	0.028	0.100	Yes
AVGO	2023–2025	-13.563	0.000***	-23.536	0.000***	0.032	0.100	Yes
AVGO	2024	-6.510	0.000***	-14.736	0.000***	0.047	0.100	Yes
AVGO	2024–2025	-11.435	0.000***	-19.341	0.000***	0.030	0.100	Yes
AVGO	2025	-12.076	0.000***	-12.080	0.000***	0.244	0.100	Yes
BOTZ	2023	-11.768	0.000***	-12.033	0.000***	0.189	0.100	Yes
BOTZ	2023–2024	-19.652	0.000***	-19.616	0.000***	0.052	0.100	Yes
BOTZ	2023–2025	-14.123	0.000***	-23.032	0.000***	0.059	0.100	Yes
BOTZ	2024	-15.874	0.000***	-16.065	0.000***	0.047	0.100	Yes
BOTZ	2024–2025	-19.575	0.000***	-19.663	0.000***	0.062	0.100	Yes

Continued on next page

Stationarity tests *(continued)*

Measure	Period	ADF_{stat}	ADF_p	PP_{stat}	PP_p	$KPSS_{stat}$	$KPSS_p$	agree
BOTZ	2025	-10.995	0.000***	-10.991	0.000***	0.117	0.100	Yes
GOOGL	2023	-13.116	0.000***	-13.268	0.000***	0.081	0.100	Yes
GOOGL	2023–2024	-20.078	0.000***	-20.329	0.000***	0.050	0.100	Yes
GOOGL	2023–2025	-13.341	0.000***	-24.053	0.000***	0.128	0.100	Yes
GOOGL	2024	-12.007	0.000***	-15.177	0.000***	0.072	0.100	Yes
GOOGL	2024–2025	-11.342	0.000***	-19.809	0.000***	0.102	0.100	Yes
GOOGL	2025	-7.490	0.000***	-12.416	0.000***	0.169	0.100	Yes
IRBO	2023	-12.947	0.000***	-13.046	0.000***	0.147	0.100	Yes
IRBO	2023–2024	-19.650	0.000***	-19.696	0.000***	0.038	0.100	Yes
IRBO	2023–2025	-13.027	0.000***	-23.527	0.000***	0.032	0.100	Yes
IRBO	2024	-14.885	0.000***	-15.367	0.000***	0.086	0.100	Yes
IRBO	2024–2025	-11.110	0.000***	-19.484	0.000***	0.060	0.100	Yes
IRBO	2025	-11.184	0.000***	-11.165	0.000***	0.175	0.100	Yes
META	2023	-13.610	0.000***	-14.130	0.000***	0.104	0.100	Yes
META	2023–2024	-16.421	0.000***	-22.190	0.000***	0.120	0.100	Yes
META	2023–2025	-24.441	0.000***	-24.547	0.000***	0.125	0.100	Yes
META	2024	-13.036	0.000***	-16.936	0.000***	0.140	0.100	Yes
META	2024–2025	-20.158	0.000***	-20.136	0.000***	0.101	0.100	Yes
META	2025	-10.972	0.000***	-10.981	0.000***	0.153	0.100	Yes
MSFT	2023	-14.987	0.000***	-15.174	0.000***	0.094	0.100	Yes
MSFT	2023–2024	-16.900	0.000***	-22.121	0.000***	0.166	0.100	Yes
MSFT	2023–2025	-24.366	0.000***	-24.441	0.000***	0.099	0.100	Yes
MSFT	2024	-12.254	0.000***	-15.940	0.000***	0.116	0.100	Yes
MSFT	2024–2025	-19.254	0.000***	-19.265	0.000***	0.088	0.100	Yes
MSFT	2025	-10.650	0.000***	-10.659	0.000***	0.286	0.100	Yes
NVDA	2023	-13.378	0.000***	-13.392	0.000***	0.158	0.100	Yes
NVDA	2023–2024	-21.878	0.000***	-21.921	0.000***	0.097	0.100	Yes
NVDA	2023–2025	-14.317	0.000***	-25.372	0.000***	0.175	0.100	Yes
NVDA	2024	-17.126	0.000***	-17.103	0.000***	0.298	0.100	Yes
NVDA	2024–2025	-11.457	0.000***	-21.153	0.000***	0.292	0.100	Yes
NVDA	2025	-6.795	0.000***	-12.324	0.000***	0.149	0.100	Yes
ROBO	2023	-11.319	0.000***	-11.535	0.000***	0.245	0.100	Yes
ROBO	2023–2024	-19.314	0.000***	-19.254	0.000***	0.044	0.100	Yes
ROBO	2023–2025	-22.517	0.000***	-22.508	0.000***	0.032	0.100	Yes
ROBO	2024	-15.802	0.000***	-16.191	0.000***	0.035	0.100	Yes
ROBO	2024–2025	-19.163	0.000***	-19.180	0.000***	0.036	0.100	Yes
ROBO	2025	-10.591	0.000***	-10.651	0.000***	0.139	0.100	Yes
TSLA	2023	-5.793	0.000***	-12.757	0.000***	0.113	0.100	Yes
TSLA	2023–2024	-20.112	0.000***	-20.164	0.000***	0.202	0.100	Yes
TSLA	2023–2025	-22.983	0.000***	-22.996	0.000***	0.059	0.100	Yes
TSLA	2024	-15.416	0.000***	-15.422	0.000***	0.403	0.076	Yes
TSLA	2024–2025	-18.902	0.000***	-18.902	0.000***	0.116	0.100	Yes
TSLA	2025	-10.834	0.000***	-10.862	0.000***	0.193	0.100	Yes
TSM	2023	-13.428	0.000***	-13.426	0.000***	0.108	0.100	Yes

Continued on next page

Stationarity tests *(continued)*

Measure	Period	ADF_{stat}	ADF_p	PP_{stat}	PP_p	$KPSS_{stat}$	$KPSS_p$	agree
TSM	2023–2024	-21.605	0.000***	-21.954	0.000***	0.091	0.100	Yes
TSM	2023–2025	-13.396	0.000***	-24.907	0.000***	0.069	0.100	Yes
TSM	2024	-16.712	0.000***	-17.305	0.000***	0.104	0.100	Yes
TSM	2024–2025	-12.341	0.000***	-20.655	0.000***	0.152	0.100	Yes
TSM	2025	-4.922	0.000***	-11.461	0.000***	0.208	0.100	Yes
VIX	2023	-14.280	0.000***	-14.676	0.000***	0.054	0.100	Yes
VIX	2023–2024	-11.478	0.000***	-22.631	0.000***	0.036	0.100	Yes
VIX	2023–2025	-23.851	0.000***	-25.569	0.000***	0.038	0.100	Yes
VIX	2024	-15.563	0.000***	-16.916	0.000***	0.024	0.100	Yes
VIX	2024–2025	-19.275	0.000***	-20.521	0.000***	0.016	0.100	Yes
VIX	2025	-11.108	0.000***	-11.175	0.000***	0.063	0.100	Yes
$EMA_c^{0.2}$	2023	-5.476	0.000***	-4.317	0.000***	0.672	0.016*	No
$EMA_c^{0.2}$	2023–2024	-7.654	0.000***	-6.643	0.000***	0.829	0.010*	No
$EMA_c^{0.2}$	2023–2025	-8.066	0.000***	-6.764	0.000***	0.763	0.010*	No
$EMA_c^{0.2}$	2024	-4.038	0.001***	-5.716	0.000***	0.484	0.045*	No
$EMA_c^{0.2}$	2024–2025	-5.756	0.000***	-5.480	0.000***	0.134	0.100	Yes
$EMA_c^{0.2}$	2025	-2.731	0.069	-2.683	0.077	0.130	0.100	No
$EMA_c^{0.8}$	2023	-11.047	0.000***	-10.464	0.000***	0.396	0.079	Yes
$EMA_c^{0.8}$	2023–2024	-17.450	0.000***	-17.485	0.000***	0.715	0.012*	No
$EMA_c^{0.8}$	2023–2025	-18.279	0.000***	-18.422	0.000***	0.588	0.024*	No
$EMA_c^{0.8}$	2024	-4.358	0.000***	-14.833	0.000***	0.315	0.100	Yes
$EMA_c^{0.8}$	2024–2025	-6.383	0.000***	-15.048	0.000***	0.091	0.100	Yes
$EMA_c^{0.8}$	2025	-6.064	0.000***	-6.168	0.000***	0.108	0.100	Yes
$AINI_c^{norm}$	2023	-13.515	0***	-13.2476	0***	0.381	0.085	Yes
$AINI_c^{norm}$	2023-2024	-21.357	0***	-21.495	0***	0.690	0.014*	No
$AINI_c^{norm}$	2023-2025	-22.532	0***	-22.884	0***	0.609	0.022*	No
$AINI_c^{norm}$	2024	-3.638	0.005**	-17.734	0.000***	0.288	0.1	Yes
$AINI_c^{norm}$	2024-2025	-6.450	0***	-18.714	0.000***	0.092	0.1	Yes
$AINI_c^{norm}$	2025	-5.583	0***	-8.034	0.000***	0.111	0.1	Yes
daily article count	2023	-6.922	0.000***	-39.083	0.000***	0.156	0.100	Yes
daily article count	2023–2024	-7.579	0.000***	-65.935	0.000***	0.128	0.100	Yes
daily article count	2023–2025	-8.339	0.000***	-70.043	0.000***	0.165	0.100	Yes
daily article count	2024	-5.622	0.000***	-38.151	0.000***	0.119	0.100	Yes
daily article count	2024–2025	-8.958	0.000***	-54.808	0.000***	0.219	0.100	Yes
daily article count	2025	-9.536	0.000***	-29.490	0.000***	0.500	0.042*	No
$EMA_{w0}^{0.2}$	2023	-4.919	0.000***	-6.152	0.000***	0.356	0.096	Yes
$EMA_{w0}^{0.2}$	2023–2024	-6.846	0.000***	-8.535	0.000***	0.301	0.100	Yes
$EMA_{w0}^{0.2}$	2023–2025	-4.813	0.000***	-7.461	0.000***	0.236	0.100	Yes
$EMA_{w0}^{0.2}$	2024	-5.395	0.000***	-5.420	0.000***	0.217	0.100	Yes
$EMA_{w0}^{0.2}$	2024–2025	-3.588	0.006**	-4.664	0.000***	0.236	0.100	Yes
$EMA_{w0}^{0.2}$	2025	-2.011	0.282	-1.924	0.321	1.129	0.010*	No
$EMA_{w0}^{0.8}$	2023	-14.726	0.000***	-14.640	0.000***	0.232	0.100	Yes
$EMA_{w0}^{0.8}$	2023–2024	-8.173	0.000***	-21.614	0.000***	0.238	0.100	Yes
$EMA_{w0}^{0.8}$	2023–2025	-7.067	0.000***	-23.989	0.000***	0.269	0.100	Yes

Continued on next page

Stationarity tests *(continued)*

Measure	Period	ADF_{stat}	ADF_p	PP_{stat}	PP_p	$KPSS_{stat}$	$KPSS_p$	agree
$EMA_{w0}^{0.8}$	2024	-14.609	0.000***	-15.284	0.000***	0.156	0.100	Yes
$EMA_{w0}^{0.8}$	2024–2025	-5.640	0.000***	-18.138	0.000***	0.268	0.100	Yes
$EMA_{w0}^{0.8}$	2025	-7.170	0.000***	-8.702	0.000***	1.240	0.010*	No
$AINI_{w0}^{norm}$	2023	-17.834	0.000***	-18.069	0.000***	0.210	0.100	Yes
$AINI_{w0}^{norm}$	2023-2024	-8.152	0.000***	-26.189	0.000***	0.241	0.100	Yes
$AINI_{w0}^{norm}$	2023-2025	-7.144	0.000***	-28.261	0.000***	0.287	0.100	Yes
$AINI_{w0}^{norm}$	2024	-18.100	0.000***	-18.500	0.000***	0.164	0.100	Yes
$AINI_{w0}^{norm}$	2024-2025	-5.980	0.000***	-21.928	0.000***	0.304	0.100	Yes
$AINI_{w0}^{norm}$	2025	-2.729	0.069	-10.802	0.000***	1.355	0.010*	No
$EMA_{w1}^{0.2}$	2023	-5.611	0.000***	-5.466	0.000***	0.215	0.100	Yes
$EMA_{w1}^{0.2}$	2023–2024	-7.754	0.000***	-7.857	0.000***	0.111	0.100	Yes
$EMA_{w1}^{0.2}$	2023–2025	-5.943	0.000***	-5.938	0.000***	0.118	0.100	Yes
$EMA_{w1}^{0.2}$	2024	-4.668	0.000***	-5.323	0.000***	0.271	0.100	Yes
$EMA_{w1}^{0.2}$	2024–2025	-3.889	0.002**	-3.929	0.002**	0.230	0.100	Yes
$EMA_{w1}^{0.2}$	2025	-2.192	0.209	-1.573	0.497	0.956	0.010*	No
$EMA_{w1}^{0.8}$	2023	-13.313	0.000***	-13.142	0.000***	0.149	0.100	Yes
$EMA_{w1}^{0.8}$	2023–2024	-20.161	0.000***	-20.754	0.000***	0.099	0.100	Yes
$EMA_{w1}^{0.8}$	2023–2025	-7.303	0.000***	-22.270	0.000***	0.139	0.100	Yes
$EMA_{w1}^{0.8}$	2024	-8.210	0.000***	-16.347	0.000***	0.238	0.100	Yes
$EMA_{w1}^{0.8}$	2024–2025	-4.870	0.000***	-16.897	0.000***	0.251	0.100	Yes
$EMA_{w1}^{0.8}$	2025	-2.733	0.069	-6.609	0.000***	0.915	0.010*	No
$AINI_{w1}^{norm}$	2023	-16.276	0.000***	-16.301	0.000***	0.134	0.100	Yes
$AINI_{w1}^{norm}$	2023-2024	-24.798	0.000***	-25.020	0.000***	0.103	0.100	Yes
$AINI_{w1}^{norm}$	2023-2025	-7.698	0.000***	-26.943	0.000***	0.151	0.100	Yes
$AINI_{w1}^{norm}$	2024	-9.110	0.000***	-19.398	0.000***	0.244	0.100	Yes
$AINI_{w1}^{norm}$	2024-2025	-5.188	0.000***	-21.061	0.000***	0.266	0.100	Yes
$AINI_{w1}^{norm}$	2025	-2.407	0.140	-8.706	0.000***	0.992	0.010*	No
$EMA_{w2}^{0.2}$	2023	-5.594	0.000***	-5.249	0.000***	0.165	0.100	Yes
$EMA_{w2}^{0.2}$	2023–2024	-7.868	0.000***	-7.945	0.000***	0.087	0.100	Yes
$EMA_{w2}^{0.2}$	2023–2025	-6.144	0.000***	-6.233	0.000***	0.183	0.100	Yes
$EMA_{w2}^{0.2}$	2024	-4.752	0.000***	-5.568	0.000***	0.163	0.100	Yes
$EMA_{w2}^{0.2}$	2024–2025	-4.014	0.001***	-4.008	0.001***	0.273	0.100	Yes
$EMA_{w2}^{0.2}$	2025	-1.441	0.563	-1.291	0.633	1.147	0.010*	No
$EMA_{w2}^{0.8}$	2023	-12.957	0.000***	-12.621	0.000***	0.105	0.100	Yes
$EMA_{w2}^{0.8}$	2023–2024	-20.219	0.000***	-20.655	0.000***	0.074	0.100	Yes
$EMA_{w2}^{0.8}$	2023–2025	-10.032	0.000***	-22.967	0.000***	0.235	0.100	Yes
$EMA_{w2}^{0.8}$	2024	-10.854	0.000***	-16.589	0.000***	0.152	0.100	Yes
$EMA_{w2}^{0.8}$	2024–2025	-5.060	0.000***	-18.247	0.000***	0.332	0.100	Yes
$EMA_{w2}^{0.8}$	2025	-2.055	0.263	-7.464	0.000***	1.136	0.010*	No
$AINI_{w2}^{norm}$	2023	-15.950	0.000***	-15.984	0.000***	0.103	0.100	Yes
$AINI_{w2}^{norm}$	2023-2024	-25.045	0.000***	-25.185	0.000***	0.070	0.100	Yes
$AINI_{w2}^{norm}$	2023-2025	-8.389	0.000***	-27.598	0.000***	0.249	0.100	Yes
$AINI_{w2}^{norm}$	2024	-12.374	0.000***	-19.780	0.000***	0.157	0.100	Yes
$AINI_{w2}^{norm}$	2024-2025	-4.317	0.000***	-22.306	0.000***	0.354	0.097	Yes

Continued on next page

Stationarity tests *(continued)*								
Measure	Period	ADF_{stat}	ADF_p	PP_{stat}	PP_p	$KPSS_{stat}$	$KPSS_p$	agree
$AINI_{w2}^{norm}$	2025	-2.222	0.198	-9.617	0.000***	1.243	0.010*	No

Table 10. Stationarity tests for all variables (rounded to three decimals). Except for raw $AINI^{var}$, tests use logged first differences of the series (returns, article counts, VIX). ADF and PP test for unit roots; KPSS tests for stationarity. The final column reports consensus across tests about stationarity (agree = stationary according to all tests). *Source:* Own calculations. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Ticker	Period	Jointly Rejected at $\alpha = 0.1$
META	2024–2025	12 (36)
IRBO	2024–2025	11 (36)
AVGO	2023–2025	8 (36)
AIQ	2024–2025	7 (36)
ROBO	2024–2025	6 (36)
NVDA	2023	6 (36)
META	2023–2025	6 (36)
TSM	2023–2025	5 (36)
ARKQ	2024	5 (36)
AVGO	2023–2024	5 (36)
AMD	2023–2024	4 (36)
AVGO	2024–2025	4 (36)
BOTZ	2024–2025	4 (36)
AAPL	2023	3 (36)
NVDA	2023–2024	3 (36)
TSM	2024–2025	3 (36)
META	2024	2 (36)
MSFT	2023	2 (36)
GOOGL	2023	2 (36)
BOTZ	2025	2 (6)
NVDA	2025	2 (6)
AVGO	2025	2 (6)
TSLA	2023	2 (36)
ARKQ	2023–2024	2 (36)
IRBO	2025	2 (6)
NVDA	2024–2025	1 (36)

Table 11. Number of Jointly Rejected Granger Causality Tests at $\alpha = 0.1$. Values in parentheses indicate total possible rejections per Ticker / Period (36 for full periods, 6 for 2025-only tests).

Number of jointly rejected H_0 hypotheses ($\alpha = 0.1$) per Ticker and Period from Granger causality tests using AINI variants as predictors of log asset returns, controlling for VIX. Total number of possible rejections given in parentheses. *Source:* Own calculations. Table generated with the assistance of generative AI (https://chatgpt.com/share/68df816e-d7b4-8013-874b-ccd146b5823b).

Granger–Causality: jointly significant results (AINI → Returns, VIX-controlled)

Model	$AINI^{\text{var}}$	Period	Ticker	β_1	ζ_1	γ_1	γ_2	γ_3	R^2_{adj}	p_a^{BH}	p_e^{BH}
w2	$EMA^{0.2}$	2023	NVDA	0.027	0.014	-0.334	0.535	-0.074	0.031	0.02**	0.02**
w2	$EMA^{0.2}$	2023	NVDA	0.021	0.013	-0.312	0.459		0.032	0.01**	0.01***
custom	$EMA^{0.2}$	2024	ARKQ	-0.060	-0.010	0.242	-0.212		0.023	0.03**	0.05**
custom	$EMA^{0.2}$	2024	META	-0.111	-0.020	-0.191			0.031	0.04**	0.07*
w2	$EMA^{0.2}$	2024-2025	AVGO	0.017	0.020	0.181	-0.726	0.511	0.024	<0.01***	0.01***
w0	$EMA^{0.2}$	2023-2024	AVGO	0.054	-0.006	0.158	-0.064	-0.321	0.014	0.01***	0.01***
w2	$EMA^{0.2}$	2024-2025	BOTZ	-0.123	-0.014	0.156	-0.222		0.015	0.04**	0.06*
w0	$EMA^{0.2}$	2023-2025	AVGO	0.018	0.021	0.148	-0.166	-0.121	0.008	0.06*	0.08*
w2	$EMA^{0.2}$	2023-2024	AMD	-0.039	-0.020	-0.146			0.001	0.05*	0.07*
w2	$EMA^{0.2}$	2024-2025	BOTZ	-0.108	-0.014	0.143	-0.337	0.143	0.021	0.08*	0.03**
w2	$EMA^{0.2}$	2024-2025	AIQ	-0.203	-0.026	0.141	-0.200		0.023	0.03**	0.06*
custom	$EMA^{0.2}$	2024-2025	NVDA	-0.137	-0.004	-0.140			0.018	0.05*	0.08*
w2	$EMA^{0.2}$	2024-2025	ROBO	-0.107	-0.012	0.139	-0.194		0.014	0.03**	0.06*
w0	$EMA^{0.2}$	2024-2025	META	-0.086	0.004	-0.134			0.011	0.04**	0.03**
w2	$EMA^{0.2}$	2024-2025	ROBO	-0.092	-0.013	0.133	-0.323	0.151	0.021	0.04**	0.03**
w2	$EMA^{0.2}$	2024-2025	AIQ	-0.183	-0.025	0.130	-0.319	0.145	0.031	0.06*	0.03**
custom	$EMA^{0.8}$	2025	NVDA	-0.128	0.058	-0.124			0.053	0.01***	0.05**
w2	$EMA^{0.8}$	2023	TSLA	0.005	-0.098	-0.122			0.031	0.03**	0.08*
custom	$AINI^{norm}$	2025	NVDA	-0.131	0.059	-0.114			0.054	0.01***	0.05**
custom	$EMA^{0.8}$	2025	AVGO	-0.075	0.101	-0.111			0.081	0.08*	0.08*
w2	$EMA^{0.2}$	2024-2025	META	-0.086	0.003	-0.109			0.011	0.02**	0.03**
w1	$EMA^{0.2}$	2024-2025	META	-0.086	0.003	-0.108			0.012	0.01***	0.02**
w2	$EMA^{0.2}$	2024-2025	IRBO	-0.100	-0.009	0.106	-0.176		0.007	0.04**	0.09*
w2	$AINI^{norm}$	2023	TSLA	0.001	-0.100	-0.100			0.030	0.03**	0.08*
custom	$AINI^{norm}$	2025	AVGO	-0.076	0.101	-0.100			0.080	0.08*	0.08*
w2	$EMA^{0.8}$	2023	NVDA	0.030	0.009	-0.097	0.092		0.021	0.05**	0.01**
w0	$EMA^{0.2}$	2023-2024	NVDA	-0.053	-0.004	-0.097	0.245	-0.380	0.005	0.09*	0.08*
custom	$EMA^{0.2}$	2024-2025	META	-0.088	0.003	-0.096			0.011	0.10*	0.05*
w2	$EMA^{0.2}$	2024-2025	IRBO	-0.085	-0.009	0.090	-0.342	0.204	0.024	0.01**	0.06*
w1	$EMA^{0.2}$	2023-2025	META	-0.060	0.003	-0.085			0.006	0.04**	0.04**
w2	$EMA^{0.2}$	2023-2025	META	-0.061	0.002	-0.084			0.005	0.06*	0.07*
w2	$EMA^{0.8}$	2023-2024	AMD	-0.035	-0.022	-0.084			0.005	0.04**	0.07*
w0	$EMA^{0.2}$	2023-2024	AMD	-0.042	-0.010	0.082	0.099	-0.430	0.007	0.09*	0.03**
w1	$EMA^{0.2}$	2024-2025	META	-0.094	0.004	-0.082	-0.036		0.008	0.01**	0.05*
w2	$EMA^{0.8}$	2023	NVDA	0.022	0.007	-0.082			-0.003	0.06*	0.07*
w2	$EMA^{0.8}$	2024-2025	AVGO	0.004	0.021	0.075	-0.130		0.012	0.00***	0.04**
w2	$AINI^{norm}$	2023	NVDA	0.031	0.008	-0.075	0.072		0.021	0.05**	0.01**
w2	$EMA^{0.2}$	2023-2025	AVGO	0.019	0.016	0.072	-0.390	0.298	0.012	0.02**	0.07*
custom	$EMA^{0.8}$	2024	ARKQ	-0.065	-0.012	0.070	-0.032		0.018	0.05*	0.05**
w2	$EMA^{0.8}$	2024-2025	AVGO	0.018	0.025	0.070	-0.131	0.019	0.012	0.01**	0.09*
w2	$AINI^{norm}$	2023	NVDA	0.023	0.005	-0.069			-0.003	0.06*	0.07*
w2	$AINI^{norm}$	2023-2024	AMD	-0.035	-0.022	-0.066			0.004	0.04**	0.07*
custom	$EMA^{0.8}$	2025	IRBO	-0.025	0.042	-0.059			0.017	0.07*	0.07*
w2	$EMA^{0.8}$	2023	GOOGL	0.031	-0.005	-0.057			0.000	0.08*	0.07*
w2	$EMA^{0.8}$	2023	MSFT	-0.103	0.004	-0.056			0.016	0.07*	0.07*
custom	$EMA^{0.8}$	2024	ARKQ	-0.051	-0.011	0.056			0.011	0.04**	0.06*
custom	$AINI^{norm}$	2024	ARKQ	-0.067	-0.012	0.054	-0.013		0.014	0.05*	0.06*
w2	$AINI^{norm}$	2024-2025	AVGO	0.007	0.021	0.054	-0.094		0.010	0.01***	0.04**
custom	$AINI^{norm}$	2025	IRBO	-0.028	0.042	-0.052			0.016	0.07*	0.07*
w1	$EMA^{0.8}$	2024-2025	META	-0.084	0.002	-0.051			0.007	0.02**	0.08*
w2	$AINI^{norm}$	2023	GOOGL	0.034	-0.006	-0.050			0.002	0.08*	0.07*
custom	$EMA^{0.8}$	2023-2024	AVGO	0.064	-0.015	-0.049			0.008	0.10*	0.09*
custom	$AINI^{norm}$	2024	ARKQ	-0.048	-0.011	0.048			0.011	0.04**	0.06*
custom	$EMA^{0.8}$	2023-2025	AVGO	0.011	0.017	-0.047			0.004	0.07*	0.07*
w2	$AINI^{norm}$	2023	MSFT	-0.100	0.003	-0.046			0.015	0.07*	0.07*
custom	$EMA^{0.8}$	2025	BOTZ	-0.091	0.008	-0.043			-0.002	0.02**	0.07*
w0	$EMA^{0.2}$	2024-2025	META	-0.092	0.005	-0.043	-0.111		0.008	0.04**	0.05**
custom	$AINI^{norm}$	2023-2024	AVGO	0.066	-0.015	-0.041			0.008	0.10*	0.09*
w1	$EMA^{0.2}$	2023-2025	META	-0.064	0.003	-0.041	-0.052		0.005	0.05**	0.05**
w1	$EMA^{0.8}$	2023-2025	META	-0.058	0.002	-0.041			0.004	0.05**	0.10*
w2	$EMA^{0.8}$	2024-2025	ROBO	-0.094	-0.011	0.041	-0.052	-0.012	0.011	0.04**	0.08*
w2	$EMA^{0.8}$	2024-2025	BOTZ	-0.108	-0.013	0.040	-0.061		0.012	0.03**	0.06*
custom	$AINI^{norm}$	2023-2025	AVGO	0.012	0.017	-0.040			0.003	0.07*	0.07*
w1	$EMA^{0.2}$	2024-2025	AIQ	-0.196	-0.024	0.039	-0.103		0.014	0.08*	0.09*
w0	$EMA^{0.8}$	2023-2024	ARKQ	0.020	-0.011	0.039	-0.027	-0.050	0.007	0.08*	0.07*
w0	$EMA^{0.8}$	2023-2024	NVDA	-0.050	-0.008	-0.038	0.041	-0.116	0.002	0.09*	0.08*
custom	$AINI^{norm}$	2025	BOTZ	-0.091	0.009	-0.038			-0.004	0.02**	0.07*
w0	$EMA^{0.8}$	2023-2025	AVGO	0.015	0.019	0.038	-0.032	-0.090	0.010	0.06*	0.06*
w2	$EMA^{0.8}$	2024-2025	ROBO	-0.095	-0.012	0.037	-0.057		0.014	0.02**	0.06*
w2	$EMA^{0.8}$	2024-2025	AIQ	-0.188	-0.025	0.037	-0.058		0.023	0.03**	0.06*
custom	$EMA^{0.2}$	2024	META	-0.141	-0.019	-0.036	-0.212		0.055	0.08*	0.03**
w0	$AINI^{norm}$	2023-2024	NVDA	-0.049	-0.008	-0.035	0.026	-0.090	0.002	0.09*	0.08*
w2	$EMA^{0.8}$	2023-2025	AVGO	0.006	0.014	0.034	-0.077		0.003	0.05*	0.09*
w2	$EMA^{0.2}$	2024-2025	TSM	-0.121	-0.017	0.034	-0.504	0.434	0.029	0.00***	0.03**
w2	$AINI^{norm}$	2024-2025	ROBO	-0.092	-0.010	0.034	-0.035	-0.019	0.011	0.04**	0.08*
w2	$EMA^{0.2}$	2023	AAPL	0.025	-0.012	-0.030	0.144		0.008	0.03**	0.04**
w2	$EMA^{0.8}$	2024-2025	IRBO	-0.099	-0.010	0.029	-0.070		0.014	0.01***	0.06*
w2	$AINI^{norm}$	2024-2025	BOTZ	-0.103	-0.012	0.029	-0.043		0.009	0.03**	0.07*
w2	$EMA^{0.8}$	2024-2025	IRBO	-0.088	-0.008	0.029	-0.066	-0.003	0.016	0.01**	0.10*
w2	$AINI^{norm}$	2024-2025	ROBO	-0.089	-0.011	0.028	-0.040		0.011	0.02**	0.06*
w2	$AINI^{norm}$	2023-2025	AVGO	0.008	0.014	0.028	-0.059		0.002	0.05*	0.09*
w2	$AINI^{norm}$	2024-2025	AIQ	-0.181	-0.024	0.027	-0.041		0.020	0.03**	0.07*
w2	$EMA^{0.8}$	2024-2025	TSM	-0.126	-0.014	0.027	-0.097		0.012	0.00***	0.07*

Continued on next page

Granger–Causality: jointly significant results (continued)

Model	$AINI^{\text{var}}$	Period	Ticker	β_1	ζ_1	γ_1	γ_2	γ_3	R^2_{adj}	p_a^{BH}	p_e^{BH}
w0	$AINI^{norm}$	2023-2024	ARKQ	0.019	-0.011	0.025	-0.020	-0.043	0.006	0.08*	0.07*
w0	$AINI^{norm}$	2023-2025	AVGO	0.016	0.019	0.025	-0.026	-0.077	0.010	0.06*	0.06*
w2	$EMA^{0.2}$	2023-2025	TSM	-0.076	-0.003	-0.025	-0.267	0.268	0.017	0.02**	0.07*
w2	$AINI^{norm}$	2024-2025	IRBO	-0.089	-0.008	0.024	-0.049	-0.016	0.017	0.01**	0.10*
w1	$EMA^{0.8}$	2024-2025	META	-0.096	0.002	-0.024	-0.054		0.010	0.01***	0.05*
w1	$EMA^{0.8}$	2023-2025	META	-0.067	0.002	-0.021	-0.046		0.008	0.01**	0.05**
w2	$AINI^{norm}$	2024-2025	IRBO	-0.092	-0.009	0.021	-0.055		0.013	0.01***	0.06*
w1	$AINI^{norm}$	2024-2025	META	-0.094	0.002	-0.020	-0.048		0.008	0.01***	0.05*
w2	$AINI^{norm}$	2024-2025	TSM	-0.121	-0.014	0.020	-0.079		0.012	0.00***	0.07*
w1	$AINI^{norm}$	2023-2025	META	-0.066	0.002	-0.020	-0.038		0.007	0.01**	0.05**
w1	$EMA^{0.8}$	2024-2025	AIQ	-0.187	-0.024	0.019	-0.051		0.017	0.08*	0.09*
w0	$EMA^{0.8}$	2024-2025	META	-0.091	0.003	-0.019	-0.070		0.007	0.04**	0.05**
w1	$AINI^{norm}$	2024-2025	AIQ	-0.182	-0.023	0.015	-0.042		0.017	0.08*	0.09*
w0	$AINI^{norm}$	2024-2025	META	-0.089	0.004	-0.014	-0.066		0.008	0.04**	0.05**
w2	$EMA^{0.8}$	2023-2025	TSM	-0.075	-0.003	-0.014	-0.070	0.036	0.014	0.02**	0.10*
w1	$EMA^{0.2}$	2024-2025	IRBO	-0.091	-0.008	-0.013	-0.169	0.105	0.014	0.06*	0.09*
w2	$EMA^{0.8}$	2023	AAPL	0.022	-0.016	-0.011	0.063	0.029	0.019	0.07*	0.06*
w2	$AINI^{norm}$	2023	AAPL	0.023	-0.016	-0.009	0.052	0.028	0.022	0.08*	0.06*
w1	$EMA^{0.8}$	2024-2025	IRBO	-0.102	-0.009	0.009	-0.064		0.014	0.02**	0.06*
w1	$AINI^{norm}$	2024-2025	IRBO	-0.091	-0.006	0.009	-0.047	-0.018	0.019	0.06*	0.09*
w0	$AINI^{norm}$	2023-2024	AVGO	0.061	-0.010	-0.008	-0.006	-0.119	0.016	0.01***	0.01***
w2	$AINI^{norm}$	2023-2025	TSM	-0.075	-0.003	-0.008	-0.056	0.016	0.012	0.02**	0.10*
w1	$AINI^{norm}$	2024-2025	IRBO	-0.096	-0.008	0.008	-0.056		0.015	0.02**	0.06*
w0	$EMA^{0.8}$	2023-2024	AVGO	0.060	-0.009	0.008	-0.005	-0.141	0.014	0.01***	0.01***
w1	$EMA^{0.8}$	2024-2025	IRBO	-0.091	-0.007	0.006	-0.057	-0.005	0.016	0.06*	0.09*
w2	$EMA^{0.8}$	2023-2025	TSM	-0.082	-0.003	-0.006	-0.061		0.007	0.02**	0.09*
w2	$AINI^{norm}$	2023-2025	TSM	-0.080	-0.003	-0.003	-0.054		0.008	0.02**	0.09*
w2	$EMA^{0.2}$	2024-2025	META	-0.095	0.003	-0.000	-0.127		0.009	0.05**	0.08*

Table 12. Granger–Causality, jointly significant results (AINI → Returns, controlled for VIX). Sorted by $|\gamma_1|$. *Source:* Own. Signif.: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Granger–Causality: jointly significant results (Returns → AINI, VIX-controlled)

Model	$AINI^{\text{var}}$	Period	Ticker	β_1	β_2	β_3	γ_1	R^2_{adj}	p_a^{BH}	p_e^{BH}
w2	$AINI^{norm}$	2023	ROBO	-0.805			0.036	0.012	0.03**	0.04**
w2	$AINI^{norm}$	2023	ROBO	-0.788	0.228	-0.971	0.042	0.020	0.00***	0.01***
w1	$AINI^{norm}$	2023	ROBO	-0.778			0.002	0.012	0.01**	0.02**
w2	$AINI^{norm}$	2023	IRBO	-0.758			0.044	0.015	0.00***	0.01***
w2	$AINI^{norm}$	2023	IRBO	-0.742	-0.115		0.041	0.004	0.02**	0.02**
w1	$AINI^{norm}$	2023	ROBO	-0.731	0.102	-0.772	-0.005	0.018	0.01***	0.03**
w1	$AINI^{norm}$	2023	ROBO	-0.715	-0.035		0.000	0.003	0.06*	0.08*
w2	$AINI^{norm}$	2023	IRBO	-0.706	-0.133	-0.664	0.037	0.011	0.01***	0.01***
w1	$AINI^{norm}$	2023	IRBO	-0.693			0.009	0.013	0.01**	0.01**
w0	$AINI^{norm}$	2023	AAPL	-0.682			-0.113	0.020	0.10*	0.09*
w1	$AINI^{norm}$	2023	IRBO	-0.670	-0.050		0.006	0.007	0.05**	0.03**
w2	$EMA^{0.8}$	2023	ROBO	-0.661			0.142	0.035	0.03**	0.04**
w2	$EMA^{0.8}$	2023	ROBO	-0.660	0.207	-0.811	0.152	0.043	0.00***	0.01***
w1	$AINI^{norm}$	2023	IRBO	-0.644	-0.053	-0.641	-0.001	0.019	0.02**	0.01**
w2	$EMA^{0.8}$	2023	IRBO	-0.629			0.151	0.039	0.00***	0.01***
w2	$EMA^{0.8}$	2023	IRBO	-0.619	-0.105		0.151	0.026	0.02**	0.02**
w1	$EMA^{0.8}$	2023	ROBO	-0.617			0.107	0.023	0.01**	0.02**
w2	$AINI^{norm}$	2023	ARKQ	-0.610			0.043	0.014	0.01**	0.02**
w2	$AINI^{norm}$	2023	ARKQ	-0.607	-0.063	-0.553	0.036	0.010	0.03**	0.04**
w2	$EMA^{0.8}$	2023	IRBO	-0.589	-0.116	-0.561	0.146	0.034	0.01***	0.01***
w2	$AINI^{norm}$	2024_25	ROBO	0.581			0.315	0.098	0.04**	0.04**
w1	$EMA^{0.8}$	2023	ROBO	-0.570	0.031	-0.566	0.105	0.024	0.01***	0.03**
w1	$EMA^{0.8}$	2023	IRBO	-0.562			0.114	0.025	0.01**	0.01**
w1	$EMA^{0.8}$	2023	ROBO	-0.558	-0.061		0.110	0.015	0.06*	0.08*
w1	$EMA^{0.8}$	2023	IRBO	-0.541	-0.085		0.116	0.020	0.05**	0.03**
w1	$AINI^{norm}$	2023	ARKQ	-0.538			0.003	0.010	0.03**	0.02**
w2	$AINI^{norm}$	2024_25	ROBO	0.525	-0.483	-0.086	0.204	0.171	0.09*	0.07*
w1	$EMA^{0.8}$	2023	IRBO	-0.519	-0.089	-0.481	0.109	0.027	0.02**	0.01**
w2	$EMA^{0.8}$	2023	ARKQ	-0.498	-0.005	-0.461	0.152	0.031	0.03**	0.04**
w2	$AINI^{norm}$	2024_25	ROBO	0.495	-0.477		0.237	0.162	0.06*	0.04**
w2	$EMA^{0.8}$	2023	ARKQ	-0.490			0.152	0.036	0.01**	0.02**
w2	$EMA^{0.8}$	2024_25	ROBO	0.488			0.483	0.230	0.04**	0.04**
w0	$EMA^{0.8}$	2023	AAPL	-0.481			-0.002	0.003	0.10*	0.10*
w2	$EMA^{0.8}$	2024_25	ROBO	0.472	-0.390	-0.011	0.347	0.283	0.09*	0.07*
w2	$EMA^{0.8}$	2024_25	ROBO	0.444	-0.399		0.378	0.278	0.05*	0.04**
w1	$EMA^{0.8}$	2023	ARKQ	-0.435			0.111	0.022	0.03**	0.02**
w2	$AINI^{norm}$	2023	AIQ	-0.393	-0.138	-0.808	0.039	-0.006	0.02**	0.08*
w2	$AINI^{norm}$	2023	AVGO	-0.360			0.053	0.010	0.07*	0.04**
w2	$AINI^{norm}$	2024_25	BOTZ	0.343	-0.483	0.008	0.205	0.169	0.08*	0.10*
w2	$AINI^{norm}$	2024_25	IRBO	0.326	-0.442		0.231	0.162	0.04**	0.09*
w2	$AINI^{norm}$	2024_25	BOTZ	0.323	-0.479		0.234	0.161	0.04**	0.06*
w2	$EMA^{0.8}$	2023	AIQ	-0.320	-0.094	-0.711	0.150	0.018	0.02**	0.08*
w1	$AINI^{norm}$	2023	AAPL	-0.312	-0.571	-0.363	-0.029	0.000	0.02**	0.08*
w2	$EMA^{0.8}$	2024_25	BOTZ	0.311	-0.387	0.057	0.349	0.281	0.08*	0.10*
w1	$AINI^{norm}$	2023	AAPL	-0.310	-0.600		-0.014	0.007	0.01**	0.04**
w2	$EMA^{0.8}$	2023	AVGO	-0.301			0.159	0.034	0.07*	0.04**
w2	$AINI^{norm}$	2023	BOTZ	-0.296	-0.123	-0.653	0.039	-0.008	0.01**	0.07*

Continued on next page

Granger–Causality: jointly significant results (continued)

Model	$AINI^{\text{var}}$	Period	Ticker	β_1	β_2	β_3	γ_1	R^2_{adj}	p_a^{BH}	p_e^{BH}
w2	$EMA^{0.8}$	2024_25	BOTZ	0.292	-0.394		0.375	0.276	0.04**	0.06*
w1	$AINI^{norm}$	2024	META	-0.291			0.168	0.045	0.04**	0.06*
w2	$AINI^{norm}$	2023	AVGO	-0.289	-0.037	-0.406	0.043	0.014	0.05*	0.02**
w2	$EMA^{0.8}$	2024_25	IRBO	0.286	-0.340		0.370	0.276	0.04**	0.09*
w1	$AINI^{norm}$	2023_24	META	-0.278			0.099	0.022	0.01**	0.03**
w1	$AINI^{norm}$	2023_24	META	-0.278	-0.074		0.089	0.020	0.04**	0.03**
w1	$AINI^{norm}$	2023_24	META	-0.272	-0.074	-0.001	0.089	0.018	0.08*	0.05**
w1	$AINI^{norm}$	2023_24_25	META	-0.257	-0.122		0.262	0.133	0.03**	0.03**
w1	$AINI^{norm}$	2023_24_25	META	-0.250			0.325	0.115	0.03**	0.03**
w2	$EMA^{0.8}$	2023	AVGO	-0.250	-0.003	-0.322	0.158	0.034	0.05*	0.02**
w1	$AINI^{norm}$	2023_24_25	META	-0.242	-0.109	-0.061	0.248	0.142	0.05*	0.06*
w1	$EMA^{0.8}$	2024	META	-0.240			0.310	0.117	0.04**	0.06*
w1	$EMA^{0.8}$	2023_24	META	-0.231	-0.064		0.211	0.062	0.04**	0.03**
w1	$EMA^{0.8}$	2023_24	META	-0.231			0.225	0.065	0.01**	0.03**
w2	$EMA^{0.8}$	2023	BOTZ	-0.229	-0.090	-0.572	0.149	0.015	0.01**	0.07*
w1	$EMA^{0.8}$	2023_24	META	-0.225	-0.063	-0.000	0.210	0.062	0.07*	0.05**
w1	$AINI^{norm}$	2023	BOTZ	-0.216	-0.207	-0.569	-0.011	-0.004	0.04**	0.10*
w1	$EMA^{0.8}$	2023_24_25	META	-0.214	-0.096		0.398	0.239	0.03**	0.03**
w1	$EMA^{0.8}$	2023_24_25	META	-0.204			0.467	0.228	0.03**	0.03**
w2	$AINI^{norm}$	2023_24	META	-0.202	-0.144	-0.126	0.052	0.009	0.02**	0.07*
w1	$EMA^{0.8}$	2023_24_25	META	-0.202	-0.086	-0.051	0.386	0.248	0.05*	0.06*
w2	$AINI^{norm}$	2023_24	META	-0.198	-0.134		0.054	0.012	0.06*	0.10*
w1	$AINI^{norm}$	2023	MSFT	0.195	-0.504		0.023	0.011	0.09*	0.10*
custom	$EMA^{0.2}$	2023_24	IRBO	0.187			0.805	0.654	0.07*	0.05*
w1	$EMA^{0.8}$	2023	AAPL	-0.186	-0.520	-0.234	0.089	0.012	0.02**	0.08*
custom	$EMA^{0.2}$	2023_24_25	ROBO	0.185			0.838	0.710	0.05**	0.04**
custom	$EMA^{0.2}$	2023_24_25	BOTZ	0.184			0.841	0.710	0.01**	0.01**
custom	$EMA^{0.2}$	2023_24_25	BOTZ	0.183	-0.064		0.954	0.711	0.03**	0.03**
w1	$EMA^{0.8}$	2023	AAPL	-0.178	-0.539		0.101	0.022	0.01**	0.04**
custom	$AINI^{norm}$	2023_24_25	AMD	0.178	0.030	0.191	0.215	0.061	0.03**	0.09*
w1	$EMA^{0.8}$	2023	MSFT	0.174	-0.423		0.137	0.027	0.09*	0.10*
custom	$EMA^{0.2}$	2023_24_25	AIQ	0.170			0.839	0.709	0.05**	0.07*
w2	$EMA^{0.8}$	2023_24	META	-0.167	-0.109	-0.106	0.175	0.042	0.02**	0.07*
custom	$EMA^{0.2}$	2023_24_25	BOTZ	0.166	-0.067	0.118	0.964	0.713	0.03**	0.02**
w2	$EMA^{0.8}$	2023_24	META	-0.163	-0.101		0.178	0.045	0.06*	0.10*
custom	$EMA^{0.2}$	2023_24_25	IRBO	0.161	-0.030		0.954	0.710	0.04**	0.08*
w1	$EMA^{0.8}$	2023	BOTZ	-0.156	-0.206	-0.428	0.101	0.005	0.04**	0.10*
w2	$EMA^{0.2}$	2024_25	ROBO	0.155			0.918	0.824	0.04**	0.04**
custom	$EMA^{0.2}$	2023_24_25	IRBO	0.155			0.840	0.710	0.02**	0.03**
custom	$EMA^{0.8}$	2023_24_25	AMD	0.148	0.009	0.170	0.385	0.159	0.03**	0.09*
w2	$EMA^{0.2}$	2024_25	ROBO	0.147	-0.087		0.875	0.828	0.05*	0.04**
w2	$EMA^{0.2}$	2024_25	ROBO	0.145	-0.087	0.018	0.873	0.826	0.09*	0.07*
w2	$AINI^{norm}$	2023_24_25	META	-0.141	-0.231		0.176	0.089	0.04**	0.05*
w2	$AINI^{norm}$	2023	TSM	-0.141	0.331	-0.348	0.076	-0.006	0.02**	0.03**
w2	$AINI^{norm}$	2023_24_25	META	-0.138	-0.229	-0.118	0.157	0.093	0.02**	0.07*
custom	$EMA^{0.2}$	2024_25	BOTZ	0.133			0.871	0.755	0.02**	0.08*
w1	$EMA^{0.2}$	2024_25	ROBO	0.132			0.928	0.847	0.08*	0.09*
w2	$EMA^{0.2}$	2023	ROBO	-0.128	0.046	-0.244	0.730	0.485	0.00***	0.05**
custom	$EMA^{0.2}$	2023_24_25	ARKQ	0.120			0.835	0.709	0.03**	0.05*
w2	$EMA^{0.2}$	2023	IRBO	-0.120	-0.043	-0.206	0.724	0.487	0.01***	0.05**
w2	$EMA^{0.8}$	2023_24_25	META	-0.116	-0.175		0.312	0.178	0.04**	0.05*
w2	$EMA^{0.8}$	2023_24_25	META	-0.115	-0.175	-0.093	0.291	0.180	0.02**	0.07*
w2	$AINI^{norm}$	2024_25	AMZN	0.112	-0.360		0.228	0.158	0.04**	0.09*
w2	$EMA^{0.2}$	2024_25	BOTZ	0.105	-0.101		0.875	0.828	0.04**	0.06*
w2	$EMA^{0.2}$	2024_25	BOTZ	0.103	-0.100	0.038	0.876	0.826	0.08*	0.10*
w2	$EMA^{0.8}$	2023	TSM	-0.098	0.259	-0.285	0.185	0.013	0.02**	0.03**
w2	$AINI^{norm}$	2024_25	META	-0.083	-0.311		0.215	0.162	0.04**	0.05**
w2	$AINI^{norm}$	2024_25	AMD	0.072	-0.226		0.226	0.164	0.07*	0.03**
w2	$EMA^{0.8}$	2024_25	META	-0.071	-0.234		0.356	0.274	0.04**	0.05**
w2	$EMA^{0.2}$	2023	AVGO	-0.068	0.032	-0.074	0.747	0.480	0.06*	0.03**
w1	$EMA^{0.2}$	2024	META	-0.067	-0.022		0.782	0.711	0.10*	0.08*
w1	$EMA^{0.2}$	2024	META	-0.065			0.839	0.713	0.04**	0.07*
w2	$EMA^{0.8}$	2024_25	AMD	0.063	-0.194		0.367	0.279	0.07*	0.03**
w1	$EMA^{0.2}$	2023_24	META	-0.063	-0.050		0.750	0.625	0.03**	0.03**
w1	$EMA^{0.2}$	2023_24	META	-0.062	-0.051	-0.004	0.748	0.625	0.06*	0.05**
w1	$EMA^{0.2}$	2023_24	META	-0.061			0.787	0.623	0.01**	0.03**
w2	$EMA^{0.8}$	2024_25	AMZN	0.060	-0.310		0.367	0.273	0.04**	0.09*
w2	$AINI^{norm}$	2023_24_25	AMD	0.058	-0.176		0.189	0.089	0.07*	0.04**
w2	$AINI^{norm}$	2023_24_25	AMD	0.057	-0.171	-0.094	0.170	0.093	0.08*	0.09*
w0	$AINI^{norm}$	2024	TSLA	-0.057	0.149	-0.045	0.125	0.021	0.01***	0.07*
w0	$AINI^{norm}$	2024	TSLA	-0.057	0.145		0.116	0.030	0.00***	0.02**
w1	$EMA^{0.2}$	2023	MSFT	0.056	-0.184		0.744	0.532	0.09*	0.04**
w2	$EMA^{0.2}$	2023	AIQ	-0.055	-0.062	-0.246	0.723	0.484	0.01**	0.08*
w2	$EMA^{0.8}$	2023_24_25	AMD	0.054	-0.142		0.324	0.179	0.07*	0.04**
w2	$AINI^{norm}$	2023_24_25	AMZN	0.054	-0.271		0.188	0.086	0.06*	0.09*
w1	$EMA^{0.2}$	2023_24_25	META	-0.054	-0.048		0.889	0.791	0.03**	0.03**
w1	$EMA^{0.2}$	2023_24_25	META	-0.053	-0.048	-0.019	0.888	0.791	0.05*	0.06*
custom	$EMA^{0.2}$	2023_24_25	AMD	0.053			0.836	0.709	0.05**	0.10*
w2	$EMA^{0.8}$	2023_24_25	AMD	0.051	-0.136	-0.066	0.304	0.181	0.08*	0.09*
w0	$EMA^{0.8}$	2024	TSLA	-0.051	0.122		0.261	0.089	0.00***	0.02**
w0	$EMA^{0.8}$	2024	TSLA	-0.050	0.124	-0.036	0.269	0.079	0.01***	0.07*
custom	$EMA^{0.2}$	2023_24_25	AMD	0.050	-0.012	0.046	0.960	0.711	0.03**	0.09*
w1	$EMA^{0.2}$	2023_24_25	META	-0.047			0.888	0.789	0.04**	0.09*
custom	$EMA^{0.2}$	2023_24_25	NVDA	0.043	0.004	0.055	0.958	0.711	0.04**	0.06*

Continued on next page

Granger–Causality: jointly significant results (continued)

Model	$AINI^{\text{var}}$	Period	Ticker	β_1	β_2	β_3	γ_1	R^2_{adj}	p_a^{BH}	p_e^{BH}
w2	$EMA^{0.2}$	2023_24	META	-0.042	-0.041	-0.035	0.728	0.570	0.02**	0.07*
w1	$EMA^{0.2}$	2023	AAPL	-0.026	-0.201		0.708	0.522	0.01**	0.04**
w2	$EMA^{0.2}$	2023_24_25	META	-0.025	-0.048	-0.032	0.832	0.752	0.02**	0.07*
w2	$EMA^{0.2}$	2023_24_25	META	-0.024	-0.046		0.840	0.754	0.05*	0.10*
w1	$EMA^{0.2}$	2023	AAPL	-0.022	-0.219	0.015	0.712	0.517	0.02**	0.07*
w0	$EMA^{0.2}$	2024	TSLA	-0.017	0.033		0.753	0.622	0.00***	0.02**
w0	$EMA^{0.2}$	2024	TSLA	-0.016	0.034	-0.007	0.757	0.618	0.01***	0.07*
w2	$EMA^{0.8}$	2023_24_25	AMZN	0.016	-0.215		0.322	0.175	0.06*	0.09*
w2	$EMA^{0.2}$	2023	TSM	-0.013	0.052	-0.092	0.753	0.475	0.05*	0.05*
w2	$EMA^{0.2}$	2024_25	AMD	0.011	-0.054		0.868	0.828	0.07*	0.03**
w2	$EMA^{0.2}$	2023_24_25	AMZN	-0.008	-0.064		0.845	0.754	0.06*	0.09*
w2	$EMA^{0.2}$	2024_25	AMZN	0.007	-0.089		0.865	0.827	0.04**	0.09*
w2	$AINI^{norm}$	2023_24	AMD	-0.004	-0.138	-0.135	0.059	0.009	0.03**	0.08*
w2	$EMA^{0.8}$	2023_24	AMD	0.003	-0.116	-0.105	0.181	0.042	0.03**	0.08*
w2	$EMA^{0.2}$	2023_24	AMD	0.000	-0.037	-0.031	0.733	0.571	0.03**	0.08*

Table 13. Granger–causality, jointly significant results (Returns → AINI, controlled for VIX). Sorted by β_i. *Source:* Own calculations. Significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

	α (total, %, compounded)	β	$t(\alpha)$	$t(\beta)$	R^2 (%)
AAPL	-20.41	1.191	-0.76	11.38	49.3
AIQ	6.55	1.323	0.50	42.64	85.3
AMD	-24.44	2.053	-0.51	17.72	41.4
AMZN	27.14	1.409	0.78	19.65	49.5
ARKQ	2.92	1.572	0.12	23.80	69.5
AVGO	118.97	2.003	1.34	14.94	40.9
BOTZ	-24.25	1.333	-1.55	30.46	74.9
GOOGL	11.83	1.124	0.31	14.07	35.1
IRBO	-23.25	1.444	-1.37	30.32	74.2
META	89.69	1.508	1.58	18.67	43.4
MSFT	16.06	1.041	0.61	15.55	48.9
NVDA	181.28	2.182	1.79	12.55	42.7
ROBO	-31.36	1.190	-2.35	38.52	75.8
TSLA	-34.45	2.325	-0.54	16.02	34.5
TSM	45.22	1.582	0.95	15.38	39.9

Table 14. Results of the CAPM regressions for all selected tickers over the full sample period (April 2023–June 2025, $N = 572$ trading days). Reported are the compounded total α (in %), the average market beta β, corresponding t-statistics, and the coefficient of determination R^2 (in %). α represents the cumulative abnormal return relative to the market benchmark, while β captures systematic exposure to market risk. *Source:* Own calculations.

Measure	**2023**	**2023–2024**	**2023–2025**	**2024**	**2024–2025**	**2025**
$EMA^{0.2}_{custom}$	5.3×10^{-5}	-5.6×10^{-5}	-1.0×10^{-6}	-1.13×10^{-4}	4.0×10^{-6}	3.53×10^{-4}
$EMA^{0.2}_{w1}$	-5.5×10^{-5}	-2.2×10^{-5}	3.0×10^{-6}	4.0×10^{-6}	3.5×10^{-5}	1.03×10^{-4}
$EMA^{0.2}_{w2}$	-6.7×10^{-5}	-7.2×10^{-5}	-2.2×10^{-5}	-7.9×10^{-5}	-3.0×10^{-6}	1.80×10^{-4}
$EMA^{0.8}_{custom}$	1.16×10^{-4}	-1.42×10^{-4}	8.7×10^{-5}	-3.22×10^{-4}	9.7×10^{-5}	1.31×10^{-3}
$EMA^{0.8}_{w0}$	-3.0×10^{-5}	-2.7×10^{-5}	-7.6×10^{-5}	-2.3×10^{-5}	-1.0×10^{-4}	-3.13×10^{-4}
$EMA^{0.8}_{w1}$	-1.4×10^{-5}	-1.49×10^{-4}	-1.70×10^{-4}	-2.45×10^{-4}	-2.52×10^{-4}	-3.00×10^{-4}
$EMA^{0.8}_{w2}$	-2.31×10^{-4}	-2.76×10^{-4}	-2.48×10^{-4}	-3.13×10^{-4}	-2.66×10^{-4}	-1.76×10^{-4}
$AINI^{norm}_{custom}$	5.0×10^{-5}	-1.86×10^{-4}	9.1×10^{-5}	-3.53×10^{-4}	1.36×10^{-4}	1.54×10^{-3}
$AINI^{norm}_{w1}$	-2.4×10^{-5}	-2.03×10^{-4}	-2.34×10^{-4}	-3.32×10^{-4}	-3.44×10^{-4}	-4.16×10^{-4}
$AINI^{norm}_{w2}$	-3.38×10^{-4}	-3.56×10^{-4}	-3.31×10^{-4}	-3.74×10^{-4}	-3.36×10^{-4}	-2.75×10^{-4}

Table 15. Covariance between AINI measures and logged growth of the VIX. *Source:* Own calculations. Table generated with the assistance of generative AI (https://chatgpt.com/share/68ee1abf-7ee8-8013-817e-7d930dda44ae).

B Data Overview

Period	AAPL	AIQ	AMD	AMZN	ARKQ	AVGO	BOTZ	GOOGL
2023	0.0846	0.1272	0.2171	0.2053	0.0731	0.3028	0.0602	0.1583
2024	0.1063	0.0857	-0.0790	0.1458	0.1158	0.2953	0.0459	0.1220
2025	-0.2038	0.0805	0.0401	-0.0134	0.0739	0.0768	-0.0190	-0.0585
2023–2024	0.0970	0.1034	0.0475	0.1712	0.0976	0.2985	0.0520	0.1375
2024–2025	0.0103	0.0841	-0.0421	0.0965	0.1028	0.2277	0.0258	0.0661
2023–2025	0.0355	0.0988	0.0460	0.1335	0.0927	0.2532	0.0375	0.0975

Period	IRBO	META	MSFT	NVDA	ROBO	TSLA	TSM	MKT
2023	0.0527	0.2728	0.1447	0.3077	0.0248	0.0960	0.0668	0.0793
2024	0.0290	0.2012	0.0482	0.3960	-0.0051	0.1927	0.2601	0.0831
2025	0.0438	0.1622	0.1168	0.0662	0.0176	-0.1810	0.0846	0.0225
2023–2024	0.0391	0.2318	0.0895	0.3583	0.0076	0.1514	0.1775	0.0815
2024–2025	0.0336	0.1891	0.0695	0.2939	0.0019	0.0770	0.2058	0.0644
2023–2025	0.0401	0.2176	0.0950	0.2986	0.0097	0.0835	0.1585	0.0694

Table 16. Average daily log returns (in %) of assets across different periods. *Source:* Own calculations.

Ticker	Year	First	Last	Yearly Growth (%)	Total Growth (%)
AAPL	2023	100.000	116.333	16.33	27.28
AAPL	2024	112.170	152.054	35.56	27.28
AAPL	2025	148.064	127.277	-14.04	27.28
AIQ	2023	100.000	127.533	27.53	79.58
AIQ	2024	124.506	158.281	27.13	79.58
AIQ	2025	157.544	179.580	13.99	79.58
AMD	2023	100.000	152.662	52.66	61.15
AMD	2024	143.517	125.093	-12.84	61.15
AMD	2025	124.928	161.154	29.00	61.15
AMZN	2023	100.000	148.364	48.36	55.34
AMZN	2024	146.402	214.227	46.33	55.34
AMZN	2025	215.038	221.023	2.78	55.34
ARKQ	2023	100.000	116.378	16.38	58.92
ARKQ	2024	113.816	155.809	36.90	58.92
ARKQ	2025	154.336	189.169	22.57	58.92
AVGO	2023	100.000	176.428	76.43	352.44
AVGO	2024	171.549	371.368	116.48	352.44
AVGO	2025	371.593	452.443	21.76	352.44
BOTZ	2023	100.000	112.165	12.17	27.73
BOTZ	2024	109.056	125.919	15.46	27.73
BOTZ	2025	127.416	127.729	0.25	27.73
GOOGL	2023	100.000	133.854	33.85	75.44
GOOGL	2024	132.397	182.049	37.50	75.44
GOOGL	2025	182.174	175.442	-3.70	75.44
IRBO	2023	100.000	110.738	10.74	34.11
IRBO	2024	108.009	119.130	10.30	34.11
IRBO	2025	119.612	134.106	12.12	34.11
META	2023	100.000	166.124	66.12	235.24
META	2024	162.524	275.847	69.73	235.24
META	2025	282.315	335.236	18.75	235.24
MSFT	2023	100.000	131.750	31.75	79.23
MSFT	2024	129.939	148.782	14.50	79.23
MSFT	2025	147.751	179.232	21.31	79.23
NVDA	2023	100.000	177.134	77.13	510.84
NVDA	2024	172.291	480.475	178.87	510.84
NVDA	2025	494.858	610.845	23.44	510.84
ROBO	2023	100.000	104.997	5.00	10.82
ROBO	2024	102.596	103.650	1.03	10.82
ROBO	2025	103.613	110.817	6.95	10.82
TSLA	2023	100.000	127.576	27.58	59.56
TSLA	2024	127.545	207.342	62.56	59.56
TSLA	2025	194.732	159.563	-18.06	59.56
TSMC	2023	100.000	113.607	13.61	164.49
TSMC	2024	110.908	218.786	97.27	164.49
TSMC	2025	223.317	264.488	18.44	164.49
S&P500	2023	100.000	115.646	15.65	51.38
S&P500	2024	114.991	142.602	24.01	51.38
S&P500	2025	142.285	151.382	6.39	51.38

Table 19. Yearly and total growth of adjusted close prices. *Note.* The table shows yearly and cumulative growth rates of rebased adjusted closing prices for all analyzed tickers between 3 Apr 2023 and 15 Jul 2025. Each series was normalized to 100 at the start of the period, enabling direct comparison of relative performance across assets. Yearly growth reflects the percentage change between the first and last trading day of each year, total growth covers the entire observation window. *Source:* Own calculations, based on data obtained via the Yahoo Finance API using the `yfinance` library (Aroussi, n.d.) on 01.08.2025. Table prepared with the assistance of generative AI (https://chatgpt.com/share/68df816e-d7b4-8013-874b-ccd146b5823b).

Descriptive Statistics of Daily Returns				
Ticker	**Period**	**Mean**	**Std. Dev.**	**n**
AAPL	2023	0.00081	0.01181	187
AAPL	2024	0.00121	0.01406	251
AAPL	2025	-0.00193	0.02643	111
AAPL	2023–2024	0.00095	0.01324	439
AAPL	2023–2025	0.00032	0.01679	551
AAPL	2024–2025	0.00018	0.01879	363
AIQ	2023	0.00130	0.01123	187
AIQ	2024	0.00096	0.01217	251
AIQ	2025	0.00071	0.02116	111
AIQ	2023–2024	0.00105	0.01181	439
AIQ	2023–2025	0.00097	0.01417	551
AIQ	2024–2025	0.00086	0.01544	363
AMD	2023	0.00226	0.02839	187
AMD	2024	-0.00055	0.03012	251
AMD	2025	-0.00034	0.03740	111
AMD	2023–2024	0.00051	0.02951	439
AMD	2023–2025	0.00034	0.03120	551
AMD	2024–2025	-0.00049	0.03243	363
AMZN	2023	0.00211	0.01943	187
AMZN	2024	0.00152	0.01771	251
AMZN	2025	-0.00034	0.02462	111
AMZN	2023–2024	0.00174	0.01844	439
AMZN	2023–2025	0.00132	0.01982	551
AMZN	2024–2025	0.00096	0.02004	363
ARKQ	2023	0.00081	0.01491	187
ARKQ	2024	0.00125	0.01647	251
ARKQ	2025	0.00065	0.02707	111
ARKQ	2023–2024	0.00101	0.01583	439
ARKQ	2023–2025	0.00092	0.01861	551
ARKQ	2024–2025	0.00104	0.02026	363
AVGO	2023	0.00304	0.02111	187
AVGO	2024	0.00308	0.03312	251
AVGO	2025	0.00065	0.03958	111
AVGO	2023–2024	0.00299	0.02859	439
AVGO	2023–2025	0.00251	0.03107	551
AVGO	2024–2025	0.00233	0.03514	363
BOTZ	2023	0.00061	0.01278	187
BOTZ	2024	0.00057	0.01344	251
BOTZ	2025	-0.00040	0.02134	111
BOTZ	2023–2024	0.00052	0.01320	439
BOTZ	2023–2025	0.00036	0.01517	551
BOTZ	2024–2025	0.00030	0.01624	363
GOOGL	2023	0.00156	0.01715	187
GOOGL	2024	0.00127	0.01765	251
GOOGL	2025	-0.00071	0.02309	111
GOOGL	2023–2024	0.00136	0.01741	439
GOOGL	2023–2025	0.00095	0.01867	551
GOOGL	2024–2025	0.00066	0.01944	363
IRBO	2023	0.00055	0.01276	187
IRBO	2024	0.00039	0.01344	251
IRBO	2025	0.00016	0.02606	111

Continued on next page

Descriptive Statistics of Daily Returns (continued)				
Ticker	**Period**	**Mean**	**Std. Dev.**	**n**
IRBO	2023–2024	0.00040	0.01318	439
IRBO	2023–2025	0.00036	0.01656	551
IRBO	2024–2025	0.00033	0.01820	363
META	2023	0.00271	0.01938	187
META	2024	0.00211	0.02248	251
META	2025	0.00118	0.02784	111
META	2023–2024	0.00231	0.02120	439
META	2023–2025	0.00212	0.02267	551
META	2024–2025	0.00188	0.02421	363
MSFT	2023	0.00147	0.01451	187
MSFT	2024	0.00054	0.01263	251
MSFT	2025	0.00117	0.01922	111
MSFT	2023–2024	0.00091	0.01346	439
MSFT	2023–2025	0.00094	0.01478	551
MSFT	2024–2025	0.00071	0.01492	363
NVDA	2023	0.00306	0.02790	187
NVDA	2024	0.00409	0.03280	251
NVDA	2025	0.00024	0.04132	111
NVDA	2023–2024	0.00358	0.03078	439
NVDA	2023–2025	0.00295	0.03315	551
NVDA	2024–2025	0.00298	0.03559	363
ROBO	2023	0.00026	0.01109	187
ROBO	2024	0.00004	0.01161	251
ROBO	2025	0.00004	0.01959	111
ROBO	2023–2024	0.00008	0.01142	439
ROBO	2023–2025	0.00007	0.01344	551
ROBO	2024–2025	0.00004	0.01448	363
TSLA	2023	0.00130	0.03004	187
TSLA	2024	0.00194	0.03945	251
TSLA	2025	-0.00138	0.04936	111
TSLA	2023–2024	0.00166	0.03566	439
TSLA	2023–2025	0.00093	0.03884	551
TSLA	2024–2025	0.00074	0.04276	363
TSM	2023	0.00068	0.01839	187
TSM	2024	0.00271	0.02533	251
TSM	2025	0.00048	0.03178	111
TSM	2023–2024	0.00178	0.02263	439
TSM	2023–2025	0.00156	0.02471	551
TSM	2024–2025	0.00208	0.02742	363

Table 20. Descriptive statistics (daily mean, standard deviation, and number of observations) for selected AI-related tickers across different sample periods. Values are based on log returns. *Source:* Own calculations based on data obtained via the Yahoo Finance API using the `yfinance` library (Aroussi, n.d.) on 01.08.2025.

Period	Measure	Lag 1	Lag 2	Lag 3
2023–2025	EMA_02_w0	0.870	0.768	0.691
2023–2025	EMA_02_w1	0.910	0.826	0.758
2023–2025	EMA_02_custom	0.884	0.748	0.624
2025	EMA_02_w0	0.930	0.870	0.824
2025	EMA_02_w1	0.957	0.906	0.865
2025	EMA_02_custom	0.923	0.809	0.679
2023–2024	EMA_02_w0	0.799	0.654	0.553
2023–2024	EMA_02_w1	0.831	0.698	0.594
2023–2024	EMA_02_custom	0.866	0.717	0.591

Table 21. Autocorrelation coefficients (γ_l) for $\text{EMA}_{0.2}$ variants across lags $l = 1, 2, 3$. All series exhibit high persistence, particularly in 2025 where $\gamma_l > 0.9$. *Source:* Own calculations.

B.1 Examples of sentiment inference

Below are three examples illustrating how FinBERT evaluates sentiment across different AI model configurations. The headline shows the original title in the WSJ, while date of publication and url (as of scraping) to the online article are given in parenthesis.

Each configuration extracts a different textual context around AI-related terms - from short local spans (*w0*) to broader contexts (*w2*) and (near) full-article coverage up to the token limit (*custom*). The corresponding hype score predicted by FinBERT is shown in parentheses.

Example 1: Energy & Utilities Roundup: Market Talk (11.10.2024; https://www.wsj.com/business/energy-oil/energy-utilities-roundup-market-talk-3991d102)

> **custom:** Energy & Utilities Roundup: Market Talk the blue chip index's free-float market-cap-weighted revenue exposure to domestic France is only 14.2%, they add. (Elena Vardon @ WSJ) advertisement 1807 ET – Woodside Energy's capex commitments prompt Macquarie to downgrade the stock to neutral from outperform. Macquarie thinks Woodside's acquisitions of LNG project developer Tellurian and the Beaumont ammonia project in the U.S. effectively extend its capex cycle to 2030. As a result, a lower dividend now looks likely, the bank adds. **(0)**
>
> **w0:** energy's capex commitments prompt. **(1)**
>
> **w1:** Woodside Energy's capex commitments prompt Macquarie to downgrade the stock to neutral from outperform. **(-1)**
>
> **w2:** ET. Woodside Energy's capex commitments prompt Macquarie to downgrade the stock to neutral from outperform. Macquarie thinks. **(0)**

Example 2: Chip Wars Boost Europe's Top Tech Company-for Now (15.01.2024; https://www.wsj.com/finance/chip-wars-boost-europes-top-tech-companyfor-now-6eb22f6e)

> **custom:** Chip Wars Boost Europe's Top Tech Company-for Now this is very challenging, particularly without access to the established supply chains in Europe, Japan and the U.S. But even if China only succeeds in reverse-engineering older lithography machines, it would close off an important revenue source for Western suppliers. Not even ASML's most advanced competitors have managed to copy its extreme-ultraviolet equipment, giving the Dutch company a monopoly on one of the underpinnings of artificial intelligence. This is one reason its stock, at 33 times forward earnings, trades at a premium to peers, helping make it Europe's most valuable technology company. But the market dominance that makes it so attractive to investors has also placed ASML squarely in the middle of the U.S.–China chip war. (–1)
>
> **w0:** Dutch company a monopoly on one of the underpinnings of artificial intelligence. **(0)**
>
> **w1:** ASML's most advanced competitors have managed to copy its extreme-ultraviolet equipment, giving the Dutch company a monopoly on one of the underpinnings of artificial intelligence. This is one reason its stock, at 33 times forward earnings, trades at a premium to peers, helping make it. **(1)**
>
> **w2:** Not even ASML's most advanced competitors have managed to copy its extreme-ultraviolet equipment, giving the Dutch company a monopoly on one of the underpinnings of artificial intelligence. This is one reason its stock, at 33 times forward earnings, trades at a premium to peers, helping make it Europe's most valuable technology company. **(1)**

Example 3: The Proud Boys Have Regrouped and Are Signaling Election Plans (04.11.2024; https://www.wsj.com/politics/national-security/the-proud-boys-have-regrouped-and-are-signaling-election-plans-de7a1f45)

> **custom:** Proud Boys Have Regrouped and Are Signaling Election Plans In one discussion-viewed by the Journal-Proud Boys members suggested "keeping your rifle by your side." In a written statement, Telegram said it "does not tolerate content that encourages disrupting legal democratic processes through violence or destruction of property." The statement added, "Telegram is aware of rising tensions during election periods and is actively ramping up its human, AI and machine-learning moderation efforts to keep the platform safe for users. We are ready to cooperate with the authorities to remove criminal content." The federal government is planning dramatically increased security for this Jan. 6 when Congress meets to certify election results. **(–1)**
>
> **w0:** AI and machine-learning moderation efforts to keep the platform safe for users. **(1)**
>
> **w1:** The statement added, "Telegram is aware of rising tensions during election periods and is actively ramping up its human, AI and machine-learning moderation efforts to keep the platform safe for users. We are ready to cooperate with the authorities to remove criminal content." **(1)**
>
> **w2:** Telegram said it "does not tolerate content that encourages disrupting legal democratic processes through violence or destruction of property." The statement added, "Telegram is aware of rising tensions during election periods and is actively ramping up its human, AI and machine-learning moderation efforts to keep the platform safe for users. We are ready to cooperate with the authorities to remove criminal content." **(0)**

C Annotation

C.1 Initial Guide

The following guide is a an exact copy of the guide submitted to the annotator, originally via PDF. The next section recites the updated guidelines: This guide is designed for annotating articles from the *The Wall Street Journal*.

Each article receives two labels:

1. **AI_RELEVANT**: 1 (yes) or 0 (no) — is the article substantially about artificial intelligence?
2. **HYPE_LEVEL** (only if AI_RELEVANT = 1):
    - 0 = Low / None
    - 1 = Moderate hype
    - 2 = High hype

Step 1: Is the article about AI?

Label as **AI_RELEVANT = 1** if:

- AI or machine learning is central to the article's topic (e.g., product launch, strategic shift, investment, regulation).
- The article covers companies, technologies, or policies explicitly involving AI.
- The focus is on AI's implications for business, markets, or society.

Label as **AI_RELEVANT = 0** if:

- AI is mentioned only in passing or as a buzzword.
- The article is about a broader topic (e.g., markets, inflation) where AI is not the core subject.

Step 2: Assign HYPE_LEVEL (only if AI_RELEVANT = 1)

0 = Low / No Hype
- Factual, technical, or skeptical tone.
- Reports slow adoption, limited impact, niche applications.
- Examples: AI research updates, infrastructure news, incremental improvements.

1 = Moderate Hype
- Describes meaningful AI use with some optimism or concern.
- Shows growing importance of AI in a business or sector.
- Balanced discussion of opportunities and risks.
- Examples: AI improving workflows, enabling automation, or being tested at scale.

2 = High Hype
- AI is framed as inevitable, transformative, or urgent.
- Language implies systemic change, large-scale disruption, or rapid adoption.
- Even if tone is neutral, the **implication** is major societal or economic impact.
- Examples: "AI Overhauls Banking Sector", "Companies Race to Embed AI", "Investors Pour Billions into AI".

Remember: FEAR can also signal hype.

- If the article warns about massive job loss, policy panic, or existential threats from AI, this also counts as high hype.

Quick Decision Guide

- Ask: "Would this article make an investor more excited (or anxious) about the significance and urgency of AI?"
- If yes → likely Level 2.
- If cautious but clear significance → Level 1.
- If minimal or background context → Level 0.

Tips

- Use the dominant message of the article.
- Headlines and lead paragraphs are most important.
- If unsure, default to Level 1.

C.2 Annotation Guide updates

Table 22. Comparison of Original and Updated Annotation Guidelines

Aspect	Original Guidelines	Updated Guidelines*
AI_RELEVANT Criteria	Strong focus on AI (e.g., strategy, products, regulation, business impact)	Same, with greater emphasis on general context
HYPE_LEVEL Options	0 = Low/None; 1 = Moderate; 2 = High	1 = Moderate; 2 = High
Level 0 (Low/None)	Factual tone; niche or slow-moving impact	Removed — all AI-relevant articles now fall under Level 1 or 2. Only non-AI = 0
Level 1 (Moderate)	Balanced tone; sector-level impact; some optimism or concern	Expanded to include any narrative of AI use
Level 2 (High)	Framed as transformative, disruptive, or urgent	Same, now explicitly includes market reactions and regulatory fears

C.3 Comparison of dictionary

Naive Label Dictionary only	Flagging Dictionary only	Intersection
Large Language Model; Large Language Models; Bias; Fine-Tuning; Prompt engineering; prompts	AI; A.I.; artificial intelligence	Machine Learning; Deep Learning
hallucination; supervised learning; unsupervised learning; autonomous systems; reinforcement learning; algorithm	GPT; ChatGPT	Neural Network; LLM
Narrow AI; speech recognition; facial recognition; Human-level AI; Artificial General Intelligence; AGI	OpenAI;	Generative AI; Transformer
social chatbots; human-robot interaction; Human-Centered Artificial Intelligence; Algorithmic bias; Artificial neural network		
Automated decision-making; Compute; Computer vision; Deepfake; Deepfakes		
Disinformation; Educational technology; Foundation model; Foundation models; Frontier AI		
Generative adversarial networks; Graphical processing units; Interpretability; Open-source; Responsible AI		
Robotics; Superintelligence; Training datasets; Transformers; TransformerArtificial Intelligence		

Table 24. Comparison of terms between the Naive Label Dictionary and the Flagging Dictionary, including their intersection. *Source:* Own.

D Declaration of Authorship

German: Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und noch nicht veröffentlicht. Ich bin mir be- wusst, dass eine unwahre Erklärung rechtliche Folgen haben wird. Ich erkläre mich damit einverstanden, dass die Arbeit mit Hilfe eines Plagiatserkennungsdienstes auf enthaltene Plagiate untersucht wird.

Lars Augustat, Recklinghausen den 20.10.2025

English: I hereby declare on my honor that I have prepared the present thesis independently and without the use of any aids other than those indicated. Thoughts and ideas taken directly or indirectly from other sources are clearly identified as such. This thesis has not been submitted to any other examination authority in the same or similar form and has not yet been published. I am aware that a false declaration will have legal consequences. I consent to this thesis being checked for plagiarism using a plagiarism detection service.

Lars Augustat, Recklinghausen, October 20, 2025