Artificial Intelligence and Generative AI in Asset Management

Large Language Models, Autonomous Agents and the Information Efficiency of Equity Markets

Roland Rupprechter[1]

October 2026

Abstract

This paper examines how generative artificial intelligence, in particular large language models (LLMs) and autonomous AI agents, changes the information efficiency of equity markets and the practice of asset management, and how it differs from the classical quantitative models of systematic investing. Building on Fama (1970) and Grossman and Stiglitz (1980), we argue that generative AI lowers the cost of processing unstructured information rather than the cost of acquiring private information, so that its first-order effect is a faster incorporation of public text into prices and a shortening of post-announcement drifts.

We review the empirical literature from the dictionary methods of Loughran and McDonald (2011) to the LLM-based return-prediction studies of Lopez-Lira and Tang (2023) and Chen, Kelly and Xiu (2022), and we set out a complete research design for an event study of AI-based sentiment measures and price reactions to earnings calls of blue-chip stocks. Applications are discussed separately for funds and for individual mandates, with documented examples from practice.

The evidence supports a sober conclusion: measurable economic benefit is so far confined to a minority of firms, spurious correlations and backtest overfitting

[1]Roland Rupprechter, Economic Research, R&B Wealth Management, Kreuzgasse 1a, 6850 Dornbirn; Member Austrian Association for Financial Analysis and Asset Management. Email: roland.rupprechter@rbvm.at. *Statement on the use of AI tools*: This paper was produced with the selective use of the AI tool Claude (Anthropic). The core elements of the content – including the research question, the objective, the methodology, the data analysis, all calculations, as well as the specialist knowledge and the layout – were produced entirely independently by the author. Under the author's guidance, Claude served as an assistant tool to support research, shorten text, verify derivations and cross-check sources. Some AI-generated text passages were thoroughly reviewed by the author and only incorporated once the content objectives had been met. Finally, the system was used for proofreading (spelling and academic style) and translation from German into English. The author takes full responsibility for the content of the paper.

remain the central methodological danger, and the EU AI Act together with MiFID II defines a governance perimeter within which the human portfolio manager retains the final decision. Generative AI strengthens, but does not replace, economically founded investment judgement.

Keywords: Artificial intelligence · Generative AI · Large language models · Market efficiency · Sentiment analysis · Event study · Earnings calls · Asset management · Fund management · Backtest overfitting · EU AI Act.

JEL codes: C45, C58, G11, G12, G14, G17, G23, O33.

Non-Technical Summary

Artificial intelligence has reached the investment industry in two waves. The first wave, which began in the 1980s, consisted of statistical and machine-learning models that search for patterns in prices, balance sheets and macroeconomic data and turn them into forecasts of returns and risks. The second wave, which began with the release of ChatGPT in November 2022, consists of generative models that read, summarise and write text and that can increasingly act on their own as so-called agents. This paper asks what the second wave changes for equity markets and for the people who manage money for others.

Our central question is whether generative AI makes stock markets more efficient, that is, whether prices absorb new information faster and more completely. The answer, drawn from the academic literature and from the first years of practice, is a qualified yes. Large language models reduce the cost of reading long and complex documents such as annual reports and earnings-call transcripts, and markets react faster to information that was always public but expensive to digest. At the same time, when many investors use similar models, their signals become more alike, crowding increases and the advantage of any single user shrinks.

To make the question tractable we describe, step by step, how a researcher would test the effect of AI-based sentiment analysis on share prices around earnings calls of large companies: which firms to study, how to measure the tone of a call with a language model, how to compute abnormal returns, and which mistakes, above all the use of models that have already seen the future, must be avoided. We also describe, in plain terms, where AI is used today in fund management and in the management of individual client mandates, from research and portfolio construction to reporting and compliance.

For investors the practical conclusion is threefold. First, AI is a powerful aid for processing information, but a survey of German asset managers shows that only a small minority has so far achieved a measurable economic return from it. Second, the more flexible a model, the easier it is to mistake chance for insight; we show that testing enough strategy variants will always produce an impressive backtest, even when none of them has any merit. Third, European regulation and the fiduciary duty of the asset manager require that a human remains accountable for every investment decision. Those who use AI as a disciplined assistant, not as an oracle, will benefit most.

Contents

Non-Technical Summary 3
1 Introduction 6
2 Conceptual Foundations: From Classical Quantitative Models to Generative AI 8
2.1 Artificial intelligence, machine learning and large language models 8
2.2 Classical AI: pattern recognition and forecasting 9
2.3 Generative AI and autonomous agents 10
2.4 Milestones of AI in finance 12
3 Information Efficiency of Equity Markets in the Age of AI 14
3.1 The efficient market hypothesis and the Grossman–Stiglitz paradox 14
3.2 Processing costs, limited attention and textual information 14
3.3 Data abundance and price informativeness 15
3.4 Hypotheses on the effect of generative AI 16
4 Empirical Evidence: Text, Sentiment and Stock Returns 17
4.1 From dictionaries to transformers 17
4.2 Earnings calls as an information event 18
4.3 Evidence from large language models 19
5 Research Design: AI-Based Sentiment Analysis and Price Reactions to Earnings Calls of Blue-Chip Stocks 22
5.1 Research question and hypotheses 22
5.2 Sample and data 22
5.3 Measuring sentiment 23
5.4 Event-study methodology 24
5.5 Expected results, pitfalls and robustness 26
6 Applications I: Fund Management 27
6.1 Signal generation and alpha research 27
6.2 Portfolio construction and risk management 28
6.3 Research and data filtering with natural language processing 28
6.4 Trading, operations and client communication 29
6.5 Practical examples 29
7 Applications II: Asset Management for Individual Mandates 31
7.1 Investment guidelines, suitability and restriction monitoring 31
7.2 Personalised reporting and client communication 32
7.3 Tax-aware optimisation and rebalancing 33
7.4 Advisor copilots and the fiduciary role of the human 33
7.5 Questions clients of individual mandates should ask 34
8 Challenges: Measurable Benefit, Spurious Correlations, Model Risk and Regulation 35

8.1 Measurable benefit versus hype 35
8.2 Spurious correlations and backtest overfitting 37
8.3 Hallucination, look-ahead bias and explainability 38
8.4 Crowding, homogenisation and systemic risk 39
8.5 The regulatory framework 40
8.6 Governance: degrees of autonomy and human oversight 42
9 Conclusion and Outlook 43
References 46

1 Introduction

Few technologies have entered asset management with as much expectation as generative artificial intelligence. Within three years of the public release of ChatGPT in November 2022, large language models (LLMs) have been embedded in the research platforms of the largest systematic managers, in the advisor desktops of global wealth managers and in the compliance functions of banks, while the first funds run largely by machine-learning systems have accumulated track records of several years.[2] The enthusiasm rests on an old promise. Since Markowitz (1952) formalised the trade-off between expected return and risk, and since Sharpe (1964) and Fama (1970) described how competitive markets impound information into prices, the investment profession has searched for systematic ways to process more information, faster and with fewer behavioural errors than human judgement allows. Generative AI appears to deliver precisely this: a reader that never tires, that digests ten thousand pages of filings in minutes and that writes a coherent summary in any language.

Yet the same promise has accompanied every previous wave of quantitative innovation, from the expert systems of the 1980s to the neural-network experiments of White (1988), the factor zoo documented by Harvey, Liu and Zhu (2016)[3] and the deep-learning asset-pricing models of Gu, Kelly and Xiu (2020). Each wave produced genuine advances, and each produced spurious discoveries, crowded trades and, occasionally, crises such as the quant meltdown of August 2007 analysed by Khandani and Lo (2011). The question this paper addresses is therefore not whether generative AI matters, but how it differs from the classical quantitative models that preceded it, and what this difference implies for the information efficiency of equity markets and for the practice of managing collective funds and individual mandates.

Our focus is the contrast between two families of models. Classical AI, in the sense of statistical learning, is built for pattern recognition and prediction: given structured inputs, it estimates a conditional expectation of returns, volatilities or default probabilities and feeds the estimate into an optimiser. Generative AI, in the sense of transformer-based language models, is built to produce content: it reads and writes, it summarises and reasons in natural language, and, when equipped with the ability to query data and other programs, it can plan and execute multi-step tasks as an agent. The economic consequences of these two capabilities are not the same. A better forecasting model raises the precision of private signals; a better reading machine lowers the cost of processing public information. In the framework of Grossman and Stiglitz (1980), the former increases the incentive to acquire

[2] Throughout this paper the term "artificial intelligence" (AI) is used in the sense of the EU Artificial Intelligence Act, Art. 3(1): a machine-based system that infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions.

[3] The expression "factor zoo" goes back to the presidential address of Cochrane (2011), who asked which of the hundreds of published return predictors are independent sources of information.

information while the latter reduces the frictions that keep public information from being fully reflected in prices, frictions that Hirshleifer and Teoh (2003) and Blankespoor, deHaan and Marinovic (2020) have described as limited attention and disclosure processing costs.

The literature on textual analysis in finance provides the empirical foundation. Tetlock (2007) showed that the tone of a newspaper column predicts short-term market movements; Loughran and McDonald (2011) demonstrated that general-purpose word lists misclassify financial language and constructed a finance-specific dictionary; Gentzkow, Kelly and Taddy (2019) and Loughran and McDonald (2016) surveyed the field as it moved from word counts to supervised learning. The transformer architecture of Vaswani et al. (2017) and the pre-trained models of Devlin et al. (2019) and Brown et al. (2020) then made it possible to measure the meaning of financial text rather than its vocabulary. Huang, Wang and Yang (2023) showed that a finance-tuned model (FinBERT) outperforms dictionaries and classical machine learning in classifying the sentiment of analyst reports and earnings calls; Lopez-Lira and Tang (2023), Chen, Kelly and Xiu (2022) and Jha et al. (2024) showed that general-purpose LLMs extract return-relevant information from headlines, news articles and earnings-call transcripts. At the same time, models trained on internet-scale corpora may in effect have read the future, which raises the question of look-ahead bias; the tests of Glasserman and Lin (2023) and He et al. (2025) suggest that this bias is smaller than feared, whereas the distraction caused by a model's prior knowledge of well-known companies can be larger.

This paper makes four contributions. First, it provides a conceptual comparison of classical and generative AI along the stages of the asset-management value chain and derives testable hypotheses about their effect on market efficiency. Second, it reviews the empirical evidence on text, sentiment and returns with particular attention to earnings calls, the most information-rich recurring event in the life of a listed company. Third, it sets out a complete research design for the example question that motivated this study, namely the influence of AI-based sentiment analysis on price reactions to earnings calls of blue-chip stocks, including sample construction, sentiment measurement, event-study econometrics and the pitfalls that distinguish a credible from a spurious result. Fourth, it documents practical applications for funds and for individual mandates and confronts them with the sobering survey evidence on measurable benefit, with the statistics of backtest overfitting and with the European regulatory framework as amended by the Digital Omnibus of July 2026. The hypotheses of Section 3.4, the research design of Section 5, the governance framework of Section 8.6 and the evaluation questions of Table 6 are the author's own proposals; they are not empirical results and are marked as such in the text.

The paper is structured as follows. Section 1 (the present section) serves as the introduction, frames the research question and surveys the prior literature. Section 2 lays the conceptual foundations and contrasts classical quantitative models with generative AI and autonomous agents. Section 3 analyses the information efficiency of equity markets in the age of AI and formulates hypotheses. Section 4 reviews the empirical evidence on text,

sentiment and returns. Section 5 develops the research design for an event study of AI-based sentiment and earnings-call reactions. Sections 6 and 7 examine applications in fund management and in the management of individual mandates. Section 8 discusses the challenges of measurable benefit, spurious correlations, model risk and regulation, and Section 9 presents the conclusion and an outlook.

2 Conceptual Foundations: From Classical Quantitative Models to Generative AI

2.1 Artificial intelligence, machine learning and large language models

The vocabulary of the field is layered. Artificial intelligence is the broadest term and denotes any system that performs tasks which would require intelligence if performed by humans. Machine learning is the subset of AI in which the rules are not programmed but estimated from data; Hastie, Tibshirani and Friedman (2009) describe its statistical core as the estimation of a function that relates inputs to outputs so as to minimise expected prediction error, and Breiman (2001) contrasted this "algorithmic modelling culture" with the data-modelling culture of classical econometrics. Deep learning, in the account of Goodfellow, Bengio and Courville (2016), is machine learning with neural networks of many layers, which learn their own representations of the data instead of relying on hand-crafted features.[4]

Large language models are deep neural networks of the transformer type introduced by Vaswani et al. (2017). They are trained on very large corpora of text to predict the next token, and in doing so they acquire a statistical representation of syntax, semantics and a substantial amount of world knowledge. Devlin et al. (2019) showed with BERT that such pre-trained representations can be adapted to specific tasks with small amounts of labelled data; Brown et al. (2020) showed with GPT-3 that sufficiently large models can perform new tasks from a handful of examples supplied in the prompt. In finance, Wu et al. (2023) trained BloombergGPT, a model of 50 billion parameters, on a corpus that combined financial documents with general text, and Huang, Wang and Yang (2023) released FinBERT, a model fine-tuned on financial communication. The term "generative AI" refers to the use of such models to produce new content, text, code, images or structured data, in contrast to the discriminative use of classifying or scoring existing content.

[4] The distinction between hand-crafted features and learned representations is central to the comparison with classical factor models, whose characteristics (value, momentum, quality, size) are economic constructs chosen by the researcher.

2.2 Classical AI: pattern recognition and forecasting

Classical AI in asset management is, at its core, applied statistics with flexible functional forms. Its inputs are structured: prices and volumes, accounting variables, analyst forecasts, macroeconomic series, and increasingly alternative data such as credit-card transactions, satellite images or job postings. Its outputs are forecasts, of expected returns, variances, correlations, default probabilities or the probability of a regime change. Mullainathan and Spiess (2017) characterised the comparative advantage of machine learning as the solution of prediction problems, as opposed to the estimation of causal parameters, and Varian (2014) and Athey and Imbens (2019) catalogued the methods that have migrated from computer science into economics: penalised regression, random forests (Breiman, 2001), gradient boosting and neural networks.[5]

The decisive empirical demonstration for equities came from Gu, Kelly and Xiu (2020), who compared these methods in predicting the cross-section of monthly US stock returns from 94 firm characteristics and found that trees and neural networks perform best, with large economic gains to investors that in some cases double the performance of leading regression-based strategies; the gains stem chiefly from non-linear interactions between characteristics. Freyberger, Neuhierl and Weber (2020) reached a similar conclusion with non-parametric methods, Bianchi, Büchner and Tamoni (2021) extended the evidence to bond risk premia, and Kelly and Xiu (2023) provided a comprehensive treatment of what they call financial machine learning. Two qualifications matter for what follows. Israel, Kelly and Moskowitz (2020) stressed that financial data are small and noisy relative to the data on which machine learning succeeds elsewhere, so that economic restrictions remain essential; Avramov, Cheng and Metzker (2023) showed that much of the apparent gain is attenuated once microcaps, distressed stocks or episodes of high market volatility are excluded.[6]

Genetic algorithms, mentioned in the brief that motivated this paper, belong to the same family. Introduced by Holland (1975), they search a large space of candidate rules or portfolio weights by selection, recombination and mutation of a population of solutions. In asset management they have been used to select trading rules, to calibrate factor weights and to solve portfolio problems with constraints that defeat convex optimisers, such as cardinality limits or round lots. Their flexibility is also their danger: an evolutionary search over thousands of rule variants is a textbook case of the multiple-testing problem that Section 8.2 analyses.

[5] Random forests average many decision trees grown on bootstrap samples of the data; gradient boosting fits trees one after another to the residuals of the previous ones. Both reduce the variance of a single tree without hand-specified functional forms.

[6] According to Avramov, Cheng and Metzker (2023), the profits of machine-learning signals stem from difficult-to-arbitrage stocks and from market states with high limits to arbitrage.

2.3 Generative AI and autonomous agents

Generative models differ from forecasting models in three respects that matter economically. First, their input is unstructured. An earnings-call transcript, a risk-factor section of an annual report, a central-bank statement or a stream of news is not a vector of numbers but a sequence of sentences whose meaning depends on context; Cao et al. (2023) showed that companies have themselves begun to write disclosures for the machines that read them. Second, their output is content rather than a point estimate. The model does not merely say that the tone of a call is negative; it can explain why, extract the specific guidance that changed, draft the note that an analyst would have written and rework it into a client letter. Third, their behaviour is instructed rather than trained: the same model performs classification, extraction, summarisation or reasoning depending on the prompt, which turns model development partly into a problem of specification and verification of instructions.[7]

Autonomous agents extend these capabilities with the ability to act. An agent is a language model that is given a goal, access to data sources and programs, and the ability to decompose the goal into steps, to call those programs, to inspect the results and to decide what to do next. In an investment context an agent can be asked to monitor a watch list, to retrieve the latest filings, to compare reported figures with consensus, to update a valuation model and to draft a recommendation with sources, all without further human prompting. Balyasny Asset Management, for instance, reports an internal research platform with specialised agents that analyse central-bank speeches or assess merger-arbitrage situations (OpenAI, 2026). The step from assistance to autonomy is also the step at which questions of accountability, model risk and regulation become acute, and Sections 7 and 8 return to it.[8]

Figure 1 contrasts the two families along the typical sequence of steps. In the classical process the model is one stage in a sequence of transformations from structured data to an optimised portfolio, and the human reviews the result at the end. In the generative process the model reads unstructured sources, produces signals and content, and an agent coordinates further steps; the human is placed in the loop before the output becomes a decision. Table 1 summarises the differences along eight dimensions.

[7] The ability to perform new tasks from instructions, called in-context learning, was documented by Brown et al. (2020); it is the property that makes a single general-purpose model usable across many investment workflows without retraining.

[8] The definition of an AI system in Art. 3(1) of the AI Act explicitly refers to "varying levels of autonomy", so that agentic systems fall within the scope of the Regulation irrespective of their underlying model.

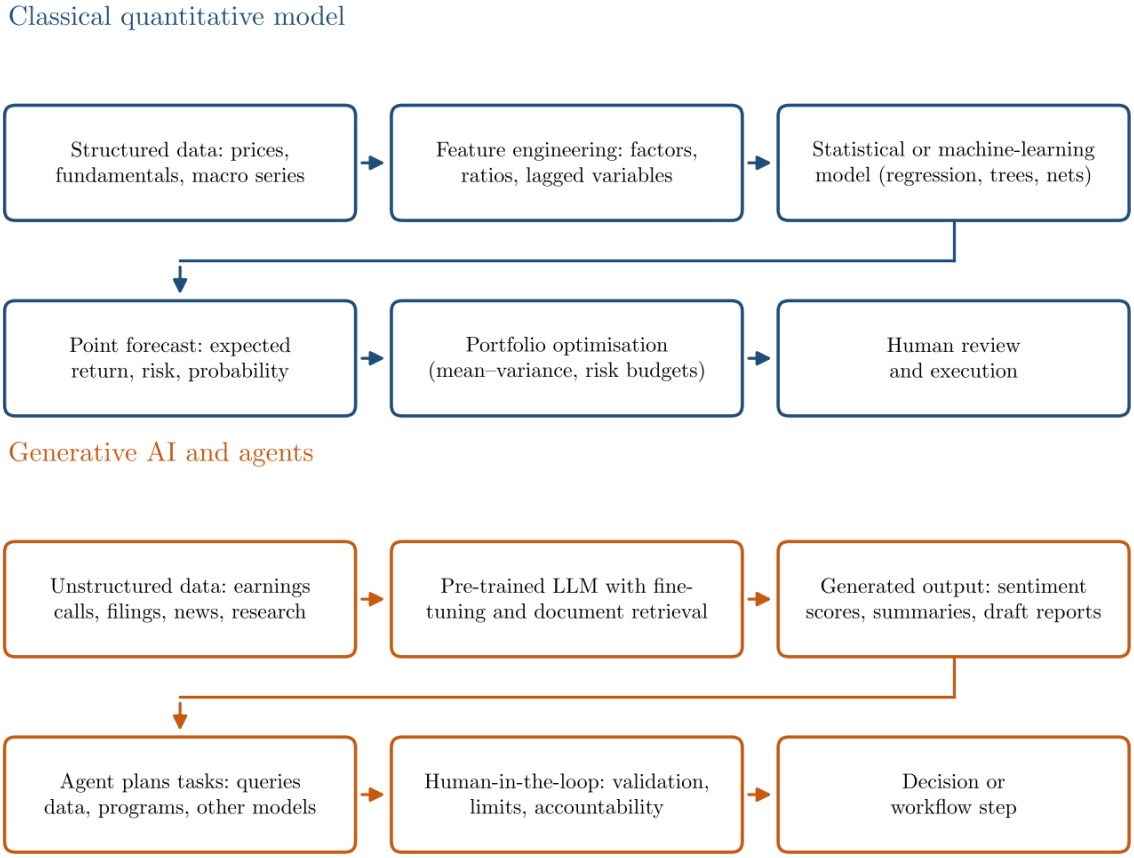


Figure 1: Classical quantitative process and generative AI process in asset management
Source: Own illustration.

Table 1: Classical AI and generative AI in asset management compared

Dimension	Classical AI / forecasting models	Generative AI / LLMs and agents
Primary purpose	Pattern recognition and prediction (returns, risk, default)	Production of content (summaries, analyses, reports) and execution of tasks
Typical models	Regression, penalised regression, trees, boosting, neural networks, genetic algorithms	Transformer-based language models (GPT, BERT, FinBERT, BloombergGPT), retrieval-augmented systems, agents
Inputs	Structured data: prices, fundamentals, macro series, alternative data	Unstructured data: transcripts, filings, news, research, e-mails; structured data via programs
Outputs	Point forecasts, probabilities, scores	Text, structured extractions, sentiment scores, draft decisions
Development	Training and validation on historical data	Pre-training by a provider; prompting, fine-tuning and retrieval by the user
Evaluation	Out-of-sample error, Sharpe ratio, calibration	Task accuracy, factual consistency, human review; look-ahead bias hard to exclude
Typical failure	Overfitting, spurious correlation, regime change	Hallucination, inconsistency, prompt sensitivity, data leakage from training corpus
Economic effect	Raises precision of private signals	Lowers cost of processing public information; accelerates workflows

Source: Own presentation, drawing on Gu, Kelly and Xiu (2020), Kelly and Xiu (2023), Brown et al. (2020) and Wu et al. (2023).

2.4 Milestones of AI in finance

The history of AI in finance is longer than the recent debate suggests, and Figure 2 places the milestones in sequence. The Dartmouth workshop of 1956 gave the field its name (McCarthy et al., 1955); the rediscovery of backpropagation by Rumelhart, Hinton and Williams (1986) made multi-layer networks trainable; White (1988) applied them to the daily returns of IBM and, with admirable candour, found no evidence against the efficient market hypothesis. Hochreiter and Schmidhuber (1997) introduced the long short-term memory network that later dominated sequence modelling. The quantitative revolution in asset management proceeded in parallel, from factor models to statistical arbitrage, and its fragility became visible in August 2007, when Khandani and Lo (2011) inferred from the simultaneous losses of many quantitative funds that their strategies had become crowded.[9]

The textual turn began with Tetlock (2007) and Loughran and McDonald (2011), the deep-learning turn with the image-recognition breakthrough of Krizhevsky, Sutskever and Hinton (2012) and the transformer of 2017. The first funds that delegated stock selection entirely to machine learning were launched in 2017, among them the Acatis AI Global

[9] Khandani and Lo (2011) reconstruct the events of 6–9 August 2007 with a simple contrarian strategy and show that losses of unprecedented size were followed by an equally rapid reversal, consistent with the forced unwinding of similar portfolios by several large funds.

Equities fund in Germany and the AI Powered Equity ETF in the United States (private banking magazin, 2017).[10] Pre-trained language models followed in 2018 and 2020, ChatGPT in 2022, BloombergGPT and the first LLM return-prediction studies in 2023. The regulatory response arrived with the EU Artificial Intelligence Act, in force since 1 August 2024, and the public statement of the European Securities and Markets Authority on AI in investment services (ESMA, 2024); the Digital Omnibus on AI of July 2026 then postponed the high-risk obligations (European Union, 2026). In the same period Bridgewater launched a fund of about two billion US dollars run largely by machine learning (Bloomberg, 2024), and agentic tools spread through the industry.

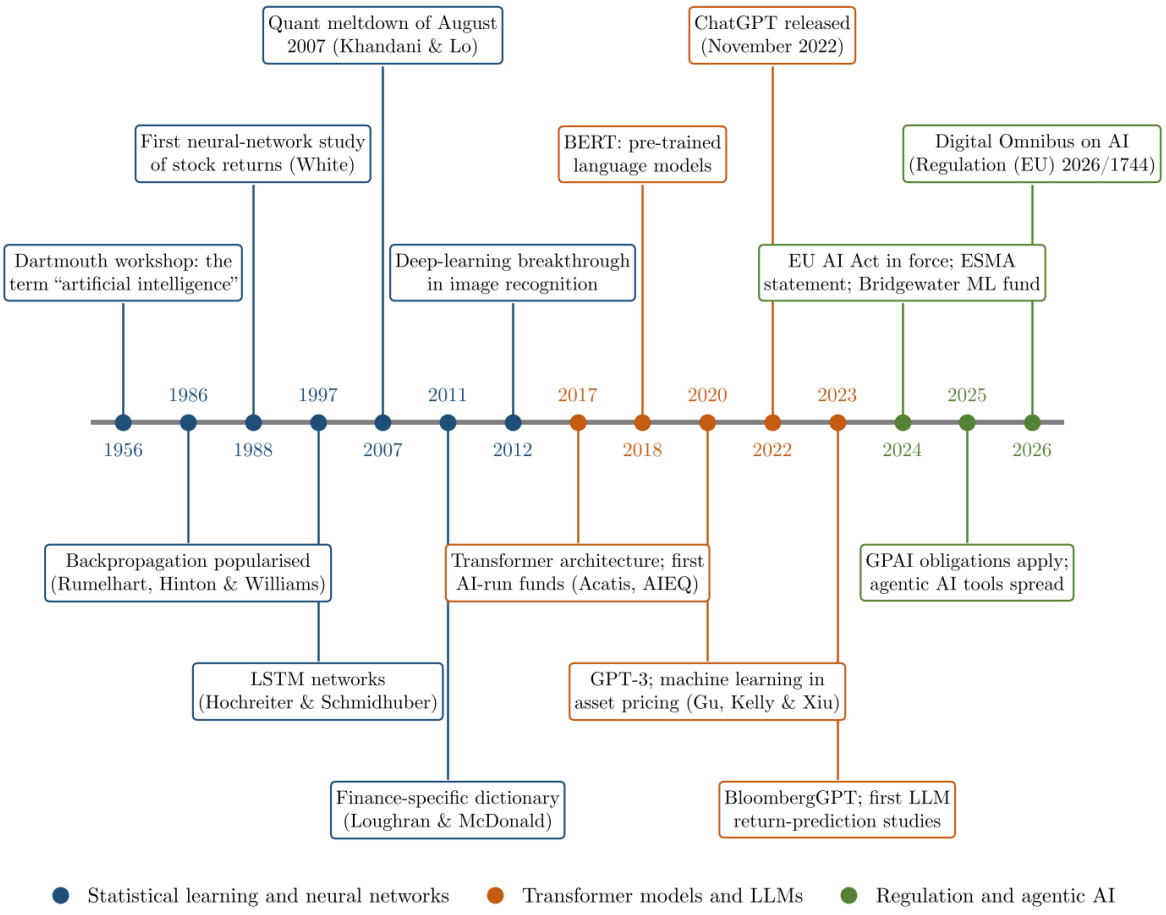


Figure 2: Milestones of artificial intelligence in finance, 1956 to 2026
Source: Own illustration based on the sources cited in the text.

[10] The AI Powered Equity ETF (ticker AIEQ) was launched in October 2017 by EquBot and uses IBM Watson to select US equities (ThinkAdvisor, 2017); according to Rupprechter (2026) it has lagged the S&P 500.

3 Information Efficiency of Equity Markets in the Age of AI

3.1 The efficient market hypothesis and the Grossman–Stiglitz paradox

Fama (1970) defined a market as efficient if prices fully reflect available information and distinguished weak, semi-strong and strong forms according to whether the information set comprises past prices, all public information or all information including private signals. The hypothesis has never been meant as a statement that prices are always right; Fama (1991) himself emphasised that efficiency can only be tested jointly with a model of expected returns. Grossman and Stiglitz (1980) supplied the reason why markets cannot be perfectly efficient: if prices revealed all information, nobody would pay to acquire it, and if nobody acquired it, prices could not reveal it. In equilibrium, informed traders earn a return on their costly information that just compensates them, and the degree of efficiency depends on the cost of information and on the noise that prevents prices from revealing it completely. Kyle (1985) formalised how informed traders exploit their advantage gradually, so that information enters prices over time rather than at once.[11]

Artificial intelligence enters this framework through the cost parameters. A technology that lowers the cost of acquiring private information raises the share of informed traders and makes prices more informative; a technology that lowers the cost of processing public information reduces the delay with which public news is reflected. The adaptive markets hypothesis of Lo (2004) adds a dynamic dimension: profit opportunities are created by innovation, exploited by early adopters and competed away as the innovation diffuses, so that efficiency is a moving target rather than a state. Stein (2009) showed that sophisticated investors can even reduce efficiency when they crowd into the same strategies and when their leverage forces them to unwind simultaneously, which is the mechanism behind the events of 2007.[12]

3.2 Processing costs, limited attention and textual information

Between the arrival of public information and its reflection in prices lies the work of reading. Hirshleifer and Teoh (2003) modelled investors with limited attention who respond to salient summary figures rather than to the detailed content of disclosures, and Da, Engelberg and Gao (2011) measured attention directly with internet search volumes. Blankespoor, deHaan and Marinovic (2020) organised the literature around the notion of disclosure processing costs: the costs of becoming aware of a disclosure, of acquiring it and

[11] Fama (1991) replaced the three forms with three families of tests, namely tests for return predictability, event studies and tests for private information; the event study of Section 5 belongs to the second family.

[12] Stein (2009) calls these the "crowded-trade" and "leverage" problems; both arise because an individual arbitrageur cannot observe how many others hold the same position, a limitation that identical AI models aggravate.

of integrating its content into a valuation. Their review argues that these costs are economically relevant, that they vary across investors and that they offer an explanation for the delayed reaction to earnings news first documented by Ball and Brown (1968) and Bernard and Thomas (1989).[13]

The most striking evidence of processing costs comes from text that is public, cheap to obtain and nevertheless ignored. Cohen, Malloy and Nguyen (2020) showed that firms which change the language of their reports subsequently underperform: there is no announcement effect, and returns accrue only when the information is later revealed through news, events or earnings; a portfolio that shorts "changers" and buys "non-changers" earned alphas of up to 188 basis points per month over 1995–2014. Fedyk and Hodson (2023) documented that recombinations of old information prompt larger price moves and subsequent reversals than direct reprints. Kim, Muhn and Nikolaev (2023) used ChatGPT to summarise corporate disclosures and found that their "bloat" is associated with higher information asymmetry among investors. These results define the territory in which generative AI can change efficiency: wherever the binding constraint is the cost of reading, a machine that reads cheaply will move prices closer to their information content.[14]

3.3 Data abundance and price informativeness

Whether more data make prices more informative is an open theoretical question. Farboodi and Veldkamp (2020) modelled the long-run growth of financial data technology and showed that extracting information from other investors' order flow crowds out the production of fundamental information; in their model this does not compromise price informativeness, but it does not necessarily improve market liquidity either. Dugast and Foucault (2018) showed that prices reflect cheap but noisy signals earlier, and that this can ultimately reduce price informativeness because it reduces the demand for more precise signals. Dessaint, Foucault and Frésard (2024) provided the empirical counterpart: with alternative data, forecasters shift their attention from the long term to the short term, so that their short-term forecasts become more and their long-term forecasts less informative. Bai, Philippon and Savov (2016) found that the informativeness of US stock prices about earnings three to five years ahead has increased since 1960, concentrated among firms with greater institutional ownership and share turnover, firms with traded options and growth firms.[15]

These results are directly relevant to generative AI. Language models are extraordinarily good at the near term: they extract what a company said yesterday and

[13] Processing costs are distinct from acquisition costs: an annual report is free to download, yet its integration into a valuation requires hours of expert reading, which is precisely the cost that language models reduce.

[14] Cohen, Malloy and Nguyen (2020) measure changes with similarity metrics between consecutive filings, such as the cosine similarity of word vectors; a fall in similarity signals that the firm has changed its language.

[15] Bai, Philippon and Savov (2016) measure price informativeness by the ability of current market valuations to predict future earnings, after controlling for current earnings.

how it differs from what it said last quarter. They are much less obviously useful for the valuation of cash flows a decade ahead, which depends on judgements about competition, technology and policy that no corpus contains. The natural hypothesis is therefore that generative AI increases short-term informational efficiency, compresses the time over which news enters prices and shortens drifts, while its effect on long-run informativeness is ambiguous and may be negative if it diverts research effort towards what can be read rather than what must be understood.[16]

3.4 Hypotheses on the effect of generative AI

Combining the arguments, we formulate four hypotheses that the research design of Section 5 and future work can test. Hypothesis 1 (speed): the share of the total price reaction to an earnings call that occurs within the first trading day has risen since the diffusion of LLM-based analysis, and the post-announcement drift has shortened. Hypothesis 2 (content): sentiment measures derived from LLMs explain a larger part of the announcement return than dictionary-based measures, and the incremental explanatory power is concentrated in qualitative passages such as the question-and-answer session rather than in the prepared remarks. Hypothesis 3 (crowding): as the same foundation models are used by many investors, the dispersion of their sentiment signals falls, the correlation of trades around announcements rises, and the profitability of a stand-alone sentiment strategy declines over time, consistent with Lo (2004) and Stein (2009). Hypothesis 4 (reflexivity): issuers adapt their language to the models, as Cao et al. (2023) documented for the machine-readability of 10-K filings, so that the measured sentiment of prepared remarks becomes less informative while the sentiment of unscripted answers remains informative.

Two further considerations shape the expected magnitude. The first is the heterogeneity of investors. Blue-chip stocks are followed by dozens of analysts and by sophisticated systematic investors, so that the marginal effect of an additional reading machine is smaller than for small caps with little coverage; the strongest effects of generative AI on efficiency should therefore appear outside the blue-chip universe, which is itself a testable implication. The second is the behaviour of the issuer. Druz et al. (2020) showed that changes in managerial tone, in particular increases in negativity ("bleak tone changes"), predict lower future earnings and that analysts and investors underreact to them; a model that measures tone surprises relative to a firm baseline is therefore, in our

[16] Dessaint, Foucault and Frésard (2024) describe this as a shift of forecasters' attention towards the short term; in our terminology it is a shortening of the investment horizon over which research adds value. The hypothesis itself is our own.

assessment, more likely to detect an effect than a model that measures tone in absolute terms. The research design in Section 5 adopts this specification.[17]

4 Empirical Evidence: Text, Sentiment and Stock Returns

4.1 From dictionaries to transformers

The measurement of sentiment in financial text has passed through three generations. The first counted words. Tetlock (2007) applied the Harvard psychosocial dictionary to the "Abreast of the Market" column of the Wall Street Journal and found that high pessimism predicts downward pressure on prices followed by a reversion, a pattern consistent with sentiment rather than information. Loughran and McDonald (2011) showed that almost three quarters of the words classified as negative by that dictionary are not negative in a financial context, "liability" and "tax" being the obvious examples, and constructed word lists for negative, positive, uncertain and litigious tone that became the standard of the field. Jegadeesh and Wu (2013) let the market itself weight the words by regressing returns on word frequencies, and García (2013) showed, with a century of New York Times columns, that the effect of sentiment on returns is concentrated in recessions.[18]

The second generation learned the weights. Ke, Kelly and Xiu (2019) proposed a supervised topic model that identifies sentiment-charged words from their association with returns and showed that it outperforms dictionaries on Dow Jones newswires. Frankel, Jennings and Lee (2022) compared machine-learning measures with dictionary-based measures at 10-K filing and conference-call dates and found a significant improvement in explanatory power for returns, with random-forest regression trees performing best. Heston and Sinha (2017) used a neural-network sentiment score and documented that daily news sentiment predicts returns for one or two days, whereas weekly aggregates predict returns for a quarter. Bybee et al. (2024) extracted topics from the Wall Street Journal and showed that the attention business news devotes to different topics closely tracks economic activity.[19]

The third generation reads. Pre-trained transformers represent each word in the context of its sentence and each sentence in the context of its document, which allows them to recognise negation, hedging and irony that word counts cannot see. Huang, Wang and Yang (2023) fine-tuned BERT on financial communication and reported that the resulting FinBERT classifies the sentiment of sentences from analyst reports more

[17] The expectation that effects concentrate in small, thinly followed stocks mirrors the evidence of Avramov, Cheng and Metzker (2023) that the gains of machine learning concentrate where limits to arbitrage are high.

[18] García (2013) uses the share of positive and negative words in two financial columns of the New York Times from 1905 to 2005; during recessions, a one standard deviation shock to his news measure predicts a change in the conditional average return of the Dow Jones Industrial Average of 12 basis points over one day.

[19] Supervised sentiment models require return labels and therefore risk learning the very noise they are meant to measure; their value can only be established out of sample.

accurately than the Loughran–McDonald dictionary and than naive Bayes, support-vector machines, random forests, convolutional and LSTM networks, with the largest advantage for small training samples and for texts containing financial words that are rare in general language; other methods underestimate the textual informativeness of earnings calls by at least 18 per cent compared with FinBERT. Wu et al. (2023) demonstrated that a model pre-trained on financial text outperforms general models of similar size on financial benchmarks while remaining competitive on general tasks. The shift from counting to reading is not merely technical: it changes the object of measurement from vocabulary to meaning, and it makes the measurement dependent on a model whose training data the researcher does not control.[20]

4.2 Earnings calls as an information event

Earnings conference calls are the most important recurring disclosure event for listed companies: management presents the quarter, gives guidance and answers questions from analysts. The call combines scripted and unscripted communication and is therefore an ideal object for textual analysis. Price et al. (2012) measured the tone of call transcripts with a finance-specific dictionary, which they found superior to the general Harvard IV-4 dictionary, and showed that call tone is a significant predictor of abnormal returns and trading volume and dominates earnings surprises over the 60 trading days after the call; the question-and-answer portion has incremental explanatory power for the post-earnings-announcement drift, concentrated in firms that do not pay dividends. Mayew and Venkatachalam (2012) went beyond words and measured the affective state of managers from the audio, showing that vocal cues are informative about future firm performance. Larcker and Zakolyukina (2012) labelled calls as deceptive on the basis of subsequent financial restatements and showed that linguistic models of the CEO's and CFO's narratives beat a random guess by 6 to 16 per cent and perform at least as well as models based on accounting variables.[21]

Druz et al. (2020) refined the measurement by showing that changes in managerial tone carry information: increases in negativity ("bleak tone changes") strongly predict lower future earnings and higher uncertainty, analysts and investors underreact to them, and a trading strategy based on them earns abnormal returns. Cao et al. (2023) documented the other side of the interaction: as machine readership of filings grew, firms avoided words perceived as negative by algorithms, and managers of firms with high machine readership even displayed speech emotion assessed as more positive and excited. The implication for

[20] FinBERT is available as an open-weight model, which allows the replication of studies without access to commercial interfaces and with a documented training corpus.

[21] In the United States, Regulation Fair Disclosure of 2000 (Securities and Exchange Commission, 2000) obliged issuers to make material information available to all investors simultaneously, which turned the earnings call into the public record on which this literature rests.

empirical work is that sentiment measures must be specified as surprises relative to a firm baseline and that the informativeness of scripted passages may decline over time as issuers learn to write for the algorithm.[22]

4.3 Evidence from large language models

The first studies with general-purpose LLMs appeared within months of the release of ChatGPT. Lopez-Lira and Tang (2023) asked the model whether a news headline was good or bad for a company's share price and formed portfolios from the answers; using headlines published after the models' knowledge cut-off, they found significant return predictability, with forecasting ability generally increasing with model size. Chen, Kelly and Xiu (2022) used LLMs such as ChatGPT and LLaMA to represent news articles from 16 equity markets in 13 languages and showed that these representations are informative about expected returns. A third study, which reported that GPT-4 predicts the direction of earnings changes from standardised financial statements better than financial analysts (Kim, Muhn and Nikolaev, 2024), was withdrawn by its authors in February 2025 after a co-author identified inconsistencies in the data and analyses; its results are therefore not relied upon here, and the episode itself illustrates how fragile early evidence in a fast-moving field can be.[23]

Beyond returns, LLMs have been used to read what quantitative models could not. Jha et al. (2024) derived from earnings-call transcripts a score of managers' anticipated changes in capital expenditure that predicts investment for up to nine quarters, and firms with high scores subsequently earn negative abnormal returns. Hansen and Kazinnik (2023) showed that GPT models classify the policy stance of FOMC announcements considerably better than other commonly used methods, and that GPT-4's explanations are on par with human reasoning. Eisfeldt, Schubert and Zhang (2023) constructed a firm-level measure of exposure to generative AI from the task content of occupations and found that highly exposed firms earned higher returns in the two weeks after the release of ChatGPT, a first indication that markets price the technology itself. Table 2 summarises the studies that are most relevant to the research design of Section 5.

A methodological caveat concerns look-ahead bias: a model whose training corpus extends beyond the sample period may "remember" how a story ended rather than infer it. Glasserman and Lin (2023) tested this by removing company names from headlines and found, surprisingly, that within the training window the anonymised headlines performed better: the general knowledge a model holds about a company distracts it more than any

[22] Cao et al. (2023) proxy machine readership with expected machine downloads of filings from the SEC's EDGAR system; mechanical downloads of 10-K and 10-Q filings rose from about 361,000 in 2003 to around 165 million in 2016.

[23] The study by Lopez-Lira and Tang (2023) has been revised several times since April 2023; its design, restricting the test to news published after the knowledge cut-off of the models, is the principle adopted in Section 5.

memorised outcome helps it, a "distraction effect" that is particularly strong for large companies, while out of sample look-ahead bias was not a concern. He et al. (2025) trained chronologically consistent models on text available up to each point in time and likewise conclude that look-ahead bias is modest in an asset-pricing application. Credible designs nonetheless exclude it by construction: they restrict the test period to dates after the model's training cut-off, anonymise entities, or use point-in-time models whose training corpus is documented. Jeff Shen of BlackRock Systematic has described the "peek-ahead bias" of off-the-shelf models as a principal reason for building finance-specific models in-house (Misonzhnik, 2025).

Table 2: Selected empirical studies on textual information, sentiment and stock returns

Study	Data and text source	Method	Main result
Tetlock (2007)	Wall Street Journal column, 1984–1999	Harvard dictionary, VAR	High pessimism predicts price pressure and reversion
Loughran and McDonald (2011)	10-K filings, 1994–2008	Finance-specific word lists	General dictionaries misclassify financial language; new lists relate to returns, volume, volatility
Price et al. (2012)	Earnings-call transcripts (US)	Finance-specific dictionary vs. Harvard IV-4	Call tone predicts abnormal returns, volume and 60-day drift; Q&A adds information for the drift
Mayew and Venkatachalam (2012)	Earnings-call audio	Vocal emotion analysis	Vocal cues of managerial affect are informative about future performance
Jegadeesh and Wu (2013)	10-K filings	Market-based word weights	Return-weighted word scores outperform equal-weighted dictionaries
Ke, Kelly and Xiu (2019)	Dow Jones newswires	Supervised sentiment topic model	Learned sentiment predicts returns better than dictionaries
Druz et al. (2020)	Earnings-call transcripts	Changes in managerial tone	Increases in negativity predict lower earnings; analysts and investors underreact
Cohen, Malloy and Nguyen (2020)	10-K and 10-Q filings, 1995–2014	Document similarity	Changes in report language predict underperformance; no announcement effect
Frankel, Jennings and Lee (2022)	10-K filings, conference calls	Machine learning (random forest best)	Learned sentiment explains returns better than dictionary measures
Huang, Wang and Yang (2023)	Analyst reports, earnings calls	FinBERT (fine-tuned transformer)	Higher classification accuracy than the LM dictionary and other ML methods
Cao et al. (2023)	10-K filings, EDGAR downloads	Machine readership measure	Firms avoid words that algorithms read as negative; more positive vocal tone
Lopez-Lira and Tang (2023)	News headlines after the models' knowledge cut-off	GPT prompts, portfolios	Significant return predictability; generally increasing with model size
Chen, Kelly and Xiu (2022)	News in 16 markets, 13 languages	LLM representations (e.g. ChatGPT, LLaMA)	LLM-based representations are informative about expected returns
Glasserman and Lin (2023)	Original vs. anonymised headlines	GPT sentiment with entity masking	In-sample, anonymised headlines do better (distraction effect); no look-ahead problem out of sample
Jha et al. (2024)	Earnings-call transcripts	ChatGPT investment score	Score predicts capital expenditure for up to nine quarters
He et al. (2025)	Point-in-time text corpora	Chronologically consistent LLMs	Look-ahead bias modest in an asset-pricing test

Source: Own presentation based on the cited studies.

5 Research Design: AI-Based Sentiment Analysis and Price Reactions to Earnings Calls of Blue-Chip Stocks

5.1 Research question and hypotheses

The example question that motivated this paper reads: what is the influence of AI-based sentiment analysis on price reactions to earnings calls, examined empirically for blue-chip stocks? The question has two layers. The first is whether sentiment measured with an LLM contains information about the price reaction beyond the quantitative earnings surprise and beyond dictionary-based tone, which is Hypothesis 2 of Section 3.4. The second is whether the diffusion of such tools among investors has changed the dynamics of the reaction itself, which is Hypothesis 1, the speed hypothesis, and Hypothesis 3, the crowding hypothesis. The design described here addresses both layers with a single data set and follows the event-study tradition of Ball and Brown (1968), Brown and Warner (1985) and MacKinlay (1997), as reviewed by Kothari and Warner (2007).

5.2 Sample and data

Blue-chip stocks are the appropriate universe for a first test for three reasons. Their transcripts are complete and available for long periods; their prices are liquid, so that abnormal returns are not contaminated by microstructure noise; and they are precisely the stocks for which one would expect the smallest effect, so that a significant result is informative. The universe comprises the constituents of the EURO STOXX 50, the DAX and the S&P 100[24] at each quarter end over the 46 quarters from the first quarter of 2015 to the second quarter of 2026. Because the DAX was enlarged from 30 to 40 members with effect from 20 September 2021 (Deutsche Börse, 2021), and German companies that belong to both the DAX and the EURO STOXX 50 are counted only once, the universe comprises around 165 companies at each quarter end up to the second quarter of 2021 and around 175 companies thereafter. On our own estimate this yields at most about 7,800 company-quarters and thus earnings calls ($26 \times 165 + 20 \times 175 = 7{,}790$); European firms that hold calls only semi-annually reduce the number further.[25] Because index memberships change, the number of distinct companies over the whole period exceeds 175; what matters for the

[24] The S&P 100 rather than the S&P 500 is used to keep the transatlantic sample balanced and to concentrate on the most heavily followed companies, for which the test is most conservative.

[25] Own calculation in four steps: (i) index places $50 + 30 + 100 = 180$ up to the second quarter of 2021 and $50 + 40 + 100 = 190$ from the third quarter of 2021; (ii) less an assumed 15 German companies that are members of both the DAX and the EURO STOXX 50 at a given quarter end, which gives about 165 and about 175 companies, while the S&P 100 does not overlap with the European indices; (iii) 26 quarters from the first quarter of 2015 to the second quarter of 2021 and 20 quarters from the third quarter of 2021 to the second quarter of 2026; (iv) $26 \times 165 + 20 \times 175 = 7{,}790$ company-quarters as an upper bound for the number of calls. The number of overlapping companies is an assumption; it varies between quarter ends with index changes and must be taken from the historical index compositions when the study is carried out.

event study is the number of events. Crucially, the sample spans a period before and a period after the diffusion of generative AI.

Three data sources are required. Transcripts of earnings calls, separated into prepared remarks and question-and-answer sessions and timestamped to the minute, are available from commercial providers such as LSEG Refinitiv or S&P Capital IQ. Daily prices, returns and market capitalisations are drawn from the same providers or from CRSP for US stocks; benchmark factors for Europe and the United States are available from the data library of Kenneth French. Analyst consensus forecasts before the call and the reported earnings per share, from which the standardised unexpected earnings are computed, are taken from I/B/E/S. The timestamp matters: many European calls take place during trading hours, and the event day must be defined as the first trading session in which the full call could have been traded on, which requires intraday data or a conservative convention.[26]

The sample is cleaned in the usual way. Calls that coincide with other major announcements, such as mergers, profit warnings or dividend cuts, are flagged and excluded in a robustness test; companies with fewer than eight prior calls are excluded so that a firm-specific tone baseline can be estimated; and estimation windows that overlap with the previous call are shortened. Table 3 summarises the design elements and their justification.

5.3 Measuring sentiment

Sentiment is measured in three ways so that the incremental contribution of generative AI can be isolated. The dictionary measure follows Loughran and McDonald (2011): net tone equals the difference between positive and negative words divided by total words, computed separately for the prepared remarks and the question-and-answer session. The fine-tuned transformer measure follows Huang, Wang and Yang (2023): each sentence is classified by FinBERT as positive, neutral or negative, and the document score is the share of positive minus the share of negative sentences. The LLM measure follows the structured prompting of Lopez-Lira and Tang (2023), extended for the present design: the model is given the anonymised transcript with company names, tickers and dates removed and is instructed to rate, on a scale from minus five to plus five, the implications for the firm's earnings over the next four quarters, to rate the confidence of management, and to list the three statements that most influenced the rating. Each transcript is scored several times at a sampling temperature of zero and the median is used.

Three precautions address the biases discussed in Section 4.3. The model must have a documented training cut-off before the beginning of the test period, or an open-weight model with a published corpus must be used, so that the test period lies outside the training data. Entity masking, as in Glasserman and Lin (2023), removes both the risk that the model recalls what happened to a company and the distraction caused by its prior

[26] The data library of Kenneth R. French provides daily factor returns for the United States and for Europe, so that abnormal returns can be computed consistently across the two regions.

knowledge of the firm, which their evidence suggests is particularly strong for large companies, the very companies of a blue-chip sample. Finally, following Druz et al. (2020), all three measures are expressed as surprises: the score of the current call minus the average score of the same firm's previous eight calls, divided by the standard deviation of those scores. The tone surprise is the variable of interest.[27]

5.4 Event-study methodology

Abnormal returns are computed relative to a factor model estimated over a window of 120 trading days ending 20 days before the call. The baseline is the market model of Brown and Warner (1985); the three-factor model of Fama and French (1993) and the four-factor model of Carhart (1997) serve as robustness checks. For firm i and event day t the abnormal return is defined by equation (1)[28]

$$\mathrm{AR}_{i,t} = \mathrm{R}_{i,t} - (\alpha_i + \beta_i \, \mathrm{R}_{m,t}) \tag{1}$$

with α_i and β_i estimated over the estimation window and $\mathrm{R}_{m,t}$ the return of the market index, and the cumulative abnormal return over the event window from τ_1 to τ_2 by equation (2)

$$\mathrm{CAR}_i(\tau_1, \tau_2) = \Sigma_{t=\tau 1}{}^{\tau 2} \, \mathrm{AR}_{i,t} \tag{2}$$

Equation (2) is evaluated for three windows: the announcement window from day 0 to day +1, the drift window from day +2 to day +30, and the full window from day 0 to day +30. Statistical significance of average abnormal returns is assessed with the standardised cross-sectional test of Boehmer, Musumeci and Poulsen (1991), which is robust to event-induced variance, and with the adjustment of Kolari and Pynnönen (2010) for cross-sectional correlation, which matters because earnings seasons cluster calls in time. Figure 3 illustrates the expected pattern in a stylised form: a jump on the event day in the direction of the sentiment surprise, followed by a drift in the same direction that is the signature of incomplete immediate reaction.

[27] The standardisation also removes differences in the scale of the three measures, so that coefficients can be compared across measures in units of one firm-specific standard deviation.

[28] The gap of 20 trading days between the end of the estimation window and the event ensures that anticipation of the announcement does not contaminate the estimated parameters.

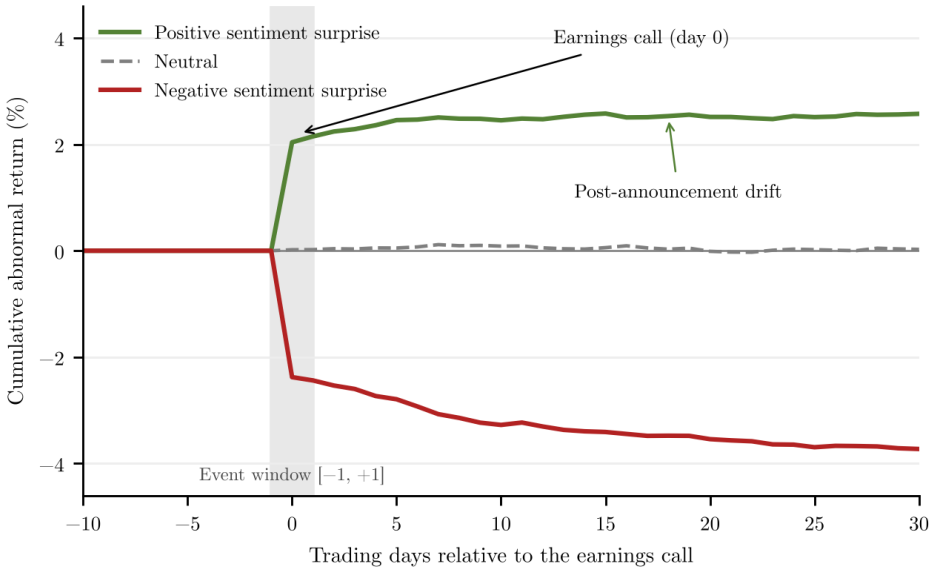


Figure 3: Stylised cumulative abnormal returns around an earnings call for positive, neutral and negative sentiment surprises
Notes: Schematic illustration of the hypothesised pattern; the paths are not estimated from data and serve only to explain the design.
Source: Own illustration.

The second stage relates the cumulative abnormal returns from equation (2) to the sentiment surprises in the cross-sectional regression of equation (3),

$$\mathrm{CAR}_i(\tau_1, \tau_2) = \gamma_0 + \gamma_1 \, \mathrm{SURP}^{\mathrm{LLM}}{}_i + \gamma_2 \, \mathrm{SURP}^{\mathrm{DICT}}{}_i + \gamma_3 \, \mathrm{SUE}_i + \gamma_4 \, \mathrm{X}_i + \varepsilon_i \qquad (3)$$

where SUE is the standardised unexpected earnings, X is a vector of controls comprising size, book-to-market, prior-year momentum, analyst coverage, a guidance-change indicator and the level of the VIX index on the event day, and the standard errors are clustered by firm and by quarter following Petersen (2009). Hypothesis 2 predicts $\gamma_1 > 0$ in equation (3), with the LLM surprise retaining significance when the dictionary surprise is included. Hypothesis 1 is tested by interacting the surprises with an indicator for calls after the first quarter of 2023 and by comparing the ratio of the announcement-window reaction to the full-window reaction before and after that date; a rising ratio and a shrinking drift coefficient, that is γ_1 in equation (3) estimated for the drift window, indicate faster incorporation. Hypothesis 3 is tested by relating the cross-sectional dispersion of LLM scores across model versions and the abnormal trading volume around calls to the diffusion indicator.[29]

[29] The diffusion of LLM-based tools among investors is not directly observable; proxies include the number of machine downloads of transcripts, survey data such as Mercer (2024) and the time elapsed since the release of ChatGPT.

Table 3: Design of the proposed event study on AI-based sentiment and earnings-call reactions

Element	Specification	Rationale and reference
Universe	EURO STOXX 50, DAX (30 members until Q2 2021, 40 thereafter) and S&P 100 constituents at each quarter end; around 165 and 175 firms	Liquid blue chips; complete transcripts; conservative test
Period	Q1 2015 to Q2 2026 (46 quarters); at most about 7,800 calls	Covers the periods before and after the diffusion of generative AI
Event	Earnings conference call, timestamped; day 0 = first full trading session	Price et al. (2012); intraday timing of European calls
Text units	Prepared remarks and Q&A session separately	Q&A adds information for the drift (Price et al., 2012)
Sentiment measures	Loughran–McDonald net tone; FinBERT sentence scores; LLM rating of anonymised transcript	Loughran and McDonald (2011); Huang, Wang and Yang (2023); Lopez-Lira and Tang (2023)
Surprise definition	Score minus firm mean of previous eight calls, standardised	Tone changes are informative (Druz et al., 2020)
Look-ahead protection	Training cut-off before test period; entity masking; point-in-time model	Glasserman and Lin (2023); He et al. (2025)
Abnormal returns	Market model; FF3 and Carhart robustness; estimation window 120 days ending t−20	Brown and Warner (1985); MacKinlay (1997)
Windows	[0, +1], [+2, +30], [0, +30]	Immediate reaction versus drift (Bernard and Thomas, 1989)
Tests	BMP standardised test; Kolari–Pynnönen adjustment; clustered regressions	Boehmer, Musumeci and Poulsen (1991); Kolari and Pynnönen (2010); Petersen (2009)
Controls	SUE, size, book-to-market, momentum, coverage, guidance change, VIX	Isolates qualitative information from the quantitative surprise
Structural change	Interaction with post-Q1 2023 indicator; drift ratio	Hypothesis 1 (speed); Lo (2004)

Source: Own presentation.

5.5 Expected results, pitfalls and robustness

The prior literature allows a reasoned, if qualitative, expectation. Price et al. (2012) and Druz et al. (2020) document economically meaningful effects of call tone on returns for broad US samples; given the higher classification accuracy reported by Huang, Wang and Yang (2023), FinBERT and LLM measures should at least match dictionary measures. For blue chips the effects will, in our assessment, be smaller than in those studies; this is an expectation, not a result. The drift coefficient, γ_1 in equation (3) estimated for the drift window, is the more interesting quantity: if Hypothesis 1 holds, the drift associated with LLM-measured surprises should have declined after 2023 even if the announcement reaction has not, which would be direct evidence that generative AI has raised the informational efficiency of blue-chip prices.

The pitfalls are those of all text-based research, sharpened by the opacity of foundation models. Multiple testing across prompts, models, windows and subsamples invites the

overfitting analysed in Section 8.2; the design therefore fixes the prompt and the windows before the data are examined and reports every specification that was run. Prompt sensitivity is addressed by repeating the scoring with paraphrased instructions and reporting the correlation of the resulting scores. Model drift is addressed by pinning the model version and archiving the raw outputs. Reverse causality, the possibility that management's tone responds to the pre-call price trend, is addressed by controlling for the pre-event run-up and by the surprise specification. Finally, the economic significance of any result must be assessed net of transaction costs and with the capacity of a strategy in mind, following the protocol of Arnott, Harvey and Markowitz (2019); a statistically significant tone effect of ten basis points in large caps is a scientific result about efficiency, not necessarily a profitable strategy.[30]

6 Applications I: Fund Management

6.1 Signal generation and alpha research

In collective funds the oldest and still most important use of AI is the generation of return forecasts, or signals, for security selection. Systematic managers combine hundreds of signals derived from prices, fundamentals, analyst behaviour and alternative data; BlackRock Systematic, with about 336 billion US dollars under management and a team of 230, reports that the number of signals in its US equity strategy grew from three at inception in 1985 to more than a thousand (Misonzhnik, 2025). Machine learning enters at two points: in the construction of individual signals from raw data, for example a sentiment score from transcripts or a demand estimate from job postings, and in the combination of signals into a forecast, where non-linear models of the kind evaluated by Gu, Kelly and Xiu (2020) replace linear weighting. Kelly, Malamud and Zhou (2024) have argued that, contrary to intuition, very high-dimensional models can improve out-of-sample performance even when they interpolate the training data, provided they are regularised appropriately; the argument supports the use of complex models but not the abandonment of economic priors.[31]

Generative AI changes the first of these points more than the second. A language model turns an earnings call into a set of structured variables, the change in guidance, the tone of the answers, the number of evasive replies, the topics that analysts pressed on, that can be fed into the same combination models as any other signal. BlackRock Systematic describes using LLMs to read alternative data, social media and regulatory filings in the

[30] Under the data policy of the author's working-paper series, all price data and every calculation step of an empirical implementation would be documented in an accompanying spreadsheet, as is done in this paper for the calculations behind Figure 5.

[31] A signal is a variable that, in the manager's model, is associated with future relative returns; the combination of many weakly informative signals is the essence of systematic investing.

role of a financial analyst and emphasises that general-purpose models must be adapted to finance to avoid peek-ahead bias (Misonzhnik, 2025). The investment horizon of such signals is short, typically a few months, which matches the near-term strength of language models discussed in Section 3.3.[32]

6.2 Portfolio construction and risk management

Portfolio construction has been the domain of optimisation since Markowitz (1952), and the practical difficulties of mean–variance optimisation, its sensitivity to estimation error in expected returns and covariances, are well known. Black and Litterman (1992) addressed the first problem by anchoring expected returns to market equilibrium and blending in investor views; AI enters naturally as a source of views, with the confidence attached to a view derived from the model's out-of-sample track record. López de Prado (2016) addressed the second problem with hierarchical risk parity, a machine-learning method that clusters assets by correlation and allocates risk along the resulting tree without inverting the covariance matrix; he reported better out-of-sample performance than minimum-variance portfolios. Cong et al. (2021) went further and trained a deep reinforcement-learning model to construct portfolios directly from firm characteristics, bypassing the forecasting step entirely.[33]

Risk management uses machine learning for the estimation of covariance structures, the detection of regime changes and the identification of crowded positions. Generative AI adds the monitoring of qualitative risks: an agent can read every new filing of every portfolio company, compare the risk-factor section with the previous version in the spirit of Cohen, Malloy and Nguyen (2020) and alert the manager to changes that a human reader would not have had time to find. The quantitative and the generative capabilities are complementary rather than competing: the former estimates how much the portfolio could lose, the latter explains what has changed in the world that could cause it.

6.3 Research and data filtering with natural language processing

The second use case named in the brief for this paper, the rapid evaluation of large volumes of unstructured data, is where generative AI has diffused fastest. A fundamental analyst covering thirty companies receives in a typical quarter thirty transcripts, several hundred broker notes, thousands of news items and a stream of regulatory filings. Language models summarise, extract and compare this material, answer questions about it with references to the source passages, and draft the first version of the analyst's note. Wellington Management reports internal tools that query its research archive, transcripts and public

[32] Misonzhnik (2025) quotes a forecast period of three to four months for BlackRock Systematic's signals and notes that the group already used natural language processing insights before the release of ChatGPT.

[33] Hierarchical risk parity requires neither the inversion nor the positive definiteness of the covariance matrix, which is why it remains stable when the number of assets exceeds the number of observations.

commentary (Correlation One, 2026), and the industry survey of Mercer (2024) found that 54 per cent of global investment managers already used AI in their investment strategy or asset-class research and a further 37 per cent planned to do so.[34]

The economic value of this use lies less in alpha than in productivity and coverage. Brynjolfsson, Li and Raymond (2025) measured an average productivity gain of about 14 to 15 per cent (depending on the version of the study) from a generative AI assistant among customer-support agents, with the largest gains for novice and low-skilled workers, Noy and Zhang (2023) found similar gains in professional writing, and Dell'Acqua et al. (2023) showed with consultants that the gains are large for tasks inside the model's competence and can turn negative outside it, which they call the jagged frontier. For research departments the implication is that generative AI widens the number of companies and documents an analyst can follow, while the quality of judgement on the difficult cases remains a human variable.

6.4 Trading, operations and client communication

In trading, machine learning has long been used to predict short-term liquidity and to choose execution strategies, and reinforcement learning is used to minimise market impact. The Flash Crash of 6 May 2010, analysed by Kirilenko et al. (2017), remains the reminder that interacting algorithms can produce dynamics that none of them intended. Dou, Goldstein and Ji (2025) showed in simulations that reinforcement-learning traders can learn to collude tacitly without communicating, which reduces price efficiency; the result echoes the pricing experiments of Calvano et al. (2020) and is one of the systemic concerns of Section 8.4. In operations, generative AI drafts fund reports, commentaries and regulatory documents, reconciles data and answers internal queries; these uses are low in glamour and high in measurable benefit.[35]

6.5 Practical examples

Table 4 collects documented examples. They span the full range from fully autonomous stock selection to assistance. At one end, the Acatis AI Global Equities fund, launched in July 2017, delegates stock selection and portfolio composition entirely to deep-learning models trained on fundamental data back to 1986, with the fund manager ensuring only implementation (private banking magazin, 2017); after almost eight years the fund managed about 37 million euros and the trade press reported an underperformance against the MSCI World over three years, which the manager related to the small- and mid-cap tilt of the model (Thieringer, 2025). Bridgewater launched in 2024 a fund of about two

[34] The analyst's note remains a human product under the fiduciary logic of Section 7.4; the generated draft shortens the time to a first version, not the responsibility for the final one.

[35] Dou, Goldstein and Ji (2025) show that two separate mechanisms underlie the collusive outcome: price-trigger strategies and a learning bias they call over-pruning; neither requires communication between the algorithms.

billion US dollars whose decisions are made largely by machine learning, including LLMs from several providers (Bloomberg, 2024). At the other end, Morgan Stanley (2024) equipped its financial advisors with OpenAI-based assistants for access to the firm's research and for meeting summaries, Goldman Sachs began rolling out its GS AI Assistant to about 10,000 employees in January 2025 with the aim of reaching all knowledge workers within the year (CNBC, 2025), and JPMorgan Chase (2024) reported more than 2,000 AI and machine-learning experts and data scientists and over 400 use cases in production. In the German-speaking market, Union Investment (2024) announced for December 2024 one of the first fund-based wealth-management offerings built on AI.[36]

Two lessons emerge from the examples. First, the fully autonomous model has existed for almost a decade and has not displaced the assisted model; the assets gathered by AI-run funds remain small relative to the industry, and their performance has been unremarkable. Second, the largest and most sophisticated managers use AI pervasively but keep humans responsible for the portfolio. Abis (2020) found differences between quantitative and discretionary mutual funds before the generative wave: quantitative funds hold considerably more stocks, trade somewhat more and specialise in stock picking, whereas discretionary funds alternate between stock picking and market timing and perform better in recessions; the performance of quantitative funds in expansions has declined over time, which she attributes to overcrowding. Generative AI may narrow the gap in adaptability, but it does not remove the need for judgement about what is novel.

[36] The author's own practice is in the assisted zone: at R&B Wealth Management, AI supports market and security research, while allocation and selection decisions are taken by the portfolio managers within the fund's investment guidelines. On the division of labour between human and machine see Rupprechter (2026).

Table 4: Documented applications of AI and generative AI in fund and asset management

Firm	Application	Year	Source
Acatis (Germany)	AI Global Equities fund: stock selection and portfolio composition entirely by deep learning on fundamentals; about 37 million euros after eight years	2017	private banking magazin (2017); Thieringer (2025)
BlackRock Systematic	More than 1,000 signals; LLMs read filings, news and alternative data; NLP signals in use before ChatGPT	since c. 2022	Misonzhnik (2025)
Bridgewater Associates	Fund of about USD 2 billion run largely by machine learning, including several LLMs	2024	Bloomberg (2024)
Balyasny Asset Management	Internal AI research platform with specialised agents for central-bank speeches and merger arbitrage	2022–2026	OpenAI (2026)
Wellington Management	Internal tools that query research archive, transcripts and public commentary	2025	Correlation One (2026)
JPMorgan Chase	More than 2,000 AI/ML experts and data scientists, over 400 use cases in production; IndexGPT trademark application for AI-based security selection	2023–2024	CNBC (2023); JPMorgan Chase (2024)
Morgan Stanley Wealth Management	OpenAI-based AI @ Morgan Stanley Assistant, fully rolled out September 2023; meeting-notes tool Debrief, June 2024	2023–2024	Morgan Stanley (2024)
Goldman Sachs	GS AI Assistant, initially for about 10,000 employees; firm-wide roll-out planned	2025	CNBC (2025)
Union Investment (Germany)	Fund-based wealth management built on AI, announced for December 2024	2024	Union Investment (2024)

Notes: Figures as stated in the sources; vendor case studies are marked by their source.
Source: Own presentation.

7 Applications II: Asset Management for Individual Mandates

7.1 Investment guidelines, suitability and restriction monitoring

An individual mandate differs from a fund in that the portfolio is tailored to one client: to the client's objectives, investment horizon, risk tolerance, tax position, liquidity needs and personal restrictions. Under Article 25 of MiFID II the manager must assess the suitability of the service for each client and must keep the portfolio within the agreed guidelines at all times (European Union, 2014). In practice these guidelines are written documents, often individually negotiated, whose rules must be converted into checks that run before and after every trade. Generative AI is unusually well suited to this task: a language model can read the mandate contract, the investment policy statement and the client's stated exclusions, extract the rules into a structured form, and an agent can apply them to the portfolio and to each proposed order, flagging breaches for human confirmation. Exclusions that depend on reading, such as a limit on revenues from a controversial activity, can be

monitored by having the model read the sustainability reports of portfolio companies, a task that previously required expensive third-party data.[37]

The same capability serves the suitability assessment itself. ESMA's statement on the use of AI in retail investment services (ESMA, 2024) reminds firms that the MiFID II obligations apply regardless of the technology employed, that the management body remains responsible for decisions supported by AI, and that firms must act in the best interest of clients, test and monitor their systems, train their staff and inform clients when AI is used in their interactions. A model that drafts the suitability report from the client questionnaire, the meeting notes and the proposed allocation reduces the administrative burden of the adviser; it does not shift the responsibility for the judgement.[38]

7.2 Personalised reporting and client communication

The most visible benefit of generative AI in individual mandates is communication. A quarterly letter that explains, in the client's language and at the client's level of expertise, what happened in the portfolio, why the manager acted as it did and what this means for the client's goals used to be a scarce good reserved for the largest mandates. Language models produce such letters from the performance attribution, the transaction list and the house view at negligible marginal cost, and they can do so for every client and in every language. Morgan Stanley (2024) reports that its advisors use an OpenAI-powered tool, Debrief, that with client consent generates meeting notes, surfaces action items and drafts a follow-up e-mail for the adviser to edit; the firm had earlier rolled out an assistant that gives advisers quick access to its research and intellectual capital.[39]

The risk in this use is factual. A letter that states a wrong return, invents a holding or misattributes a loss damages trust irreparably, and the tendency of language models to produce fluent but false statements, documented by Ji et al. (2023), is not a rare malfunction but a property of the technology. The remedy is architectural: every number in a generated letter must be drawn from the verified reporting data base and inserted, not generated, and every qualitative statement must be traceable to a source passage. Firms that follow this discipline obtain reports that are more accurate than hand-written ones; firms that let the model write freely obtain elegant fiction.

[37] The extraction of rules from contracts is an information-extraction task in which transformer models reach high accuracy; nevertheless every extracted rule must be confirmed by the responsible person before it is enforced.

[38] ESMA (2024) is addressed to firms providing investment services to retail clients and explicitly includes the use of third-party AI tools by staff; it is a reminder of existing obligations, not a new rule.

[39] Periodic reporting to portfolio-management clients is mandatory at least every three months under Article 60 of Commission Delegated Regulation (EU) 2017/565 (European Union, 2017), which gives the generated letter a regulatory as well as a commercial function.

7.3 Tax-aware optimisation and rebalancing

Individual mandates allow optimisation at the level of tax lots, which collective funds cannot offer. Rebalancing with regard to realised gains and losses, the harvesting of losses within the limits of national tax rules and the construction of personalised index portfolios, known as direct indexing, are computational problems that machine learning handles well: the objective combines tracking error, expected return, transaction costs and tax liabilities, and the constraints are specific to each client. Here classical optimisation rather than generative AI does the work, but the generative layer makes the result usable: it explains to the client, and to the adviser, why the proposed trades were chosen and what the alternatives would have cost.[40]

7.4 Advisor copilots and the fiduciary role of the human

The adviser who manages an individual mandate performs two functions that machines do not: the exercise of fiduciary judgement and the management of the client's behaviour. The first is a legal and ethical position; the manager is accountable for acting in the client's best interest, and neither MiFID II nor the AI Act allows this accountability to be delegated to a system. The second is behavioural. Kahneman (2011) and the literature on investor behaviour that Barberis, Shleifer and Vishny (1998) and Daniel, Hirshleifer and Subrahmanyam (1998) formalised describe investors who overreact to news, extrapolate recent returns and abandon strategies at the worst moment; in the author's previous paper (Rupprechter, 2026) the capacity to hold a client to a written strategy through a drawdown was identified as one of the areas in which the human adviser retains a clear advantage. An AI copilot that prepares the facts, drafts the scenario analysis and documents the conversation strengthens this function; a chatbot that answers the client's anxious question at midnight with a fluent but ungrounded reassurance undermines it.[41]

The appropriate division of labour follows the logic of Figure 7 in Section 8.6: the lower the consequence of an error and the easier its detection, the more autonomy the system may have. Summarising a meeting is at one end; deciding to liquidate a retiree's equity allocation is at the other. Table 5 sets out a stylised workflow for an individual mandate with the role of AI and the role of the human at each step.

[40] In Austria the custodian offsets losses and gains across all securities accounts the taxpayer holds with it (§ 93 Abs. 6 EStG 1988; business, trust and jointly held accounts are excluded); offsetting across banks requires the tax return. Any tax-aware optimisation must therefore model the actual custodial structure of the client.

[41] Daniel, Hirshleifer and Subrahmanyam (1998) attribute overreaction to overconfidence and self-attribution bias; both are amplified when a fluent machine confirms the investor's prior belief.

Table 5: Stylised workflow of an individual mandate with the roles of AI and of the human

Step	Role of AI	Role of the human	Control
Onboarding and suitability	Extracts client data from documents; drafts suitability assessment and investment policy statement	Conducts the conversation; judges objectives and risk capacity; signs the assessment	Four-eyes review; MiFID II record keeping
Strategic allocation	Scenario analysis; capital-market assumptions; draft proposal	Sets the allocation within house view and client constraints	Investment committee
Security selection	Reads filings and calls; produces sentiment and quality scores; screens exclusions	Selects securities; overrides with documented reasons	Model validation; override log
Order generation	Proposes rebalancing trades with tax and cost optimisation	Approves orders	Pre-trade guideline check
Monitoring	Continuous guideline and limit monitoring; reads news on holdings; alerts	Decides on alerts	Post-trade compliance; escalation rules
Reporting	Drafts personalised letters from verified data; summarises meetings	Reviews and signs; explains in person	Grounding of all figures; audit trail
Review	Compares outcome with objectives; flags drift from strategy	Holds the client to the strategy; revises if circumstances change	Annual suitability review

Source: Own presentation.

7.5 Questions clients of individual mandates should ask

For clients the question is not whether their manager uses AI, which within a few years all managers will, but how. Table 6 lists the questions a client should ask and the warning signs in the answers. The questions apply equally to the investor who considers a fund that advertises its use of AI: the existence of a model is no evidence of its value, and a backtest that was not pre-registered, that reports no transaction costs and that cannot be explained in economic terms should be treated as the result of a search rather than as a discovery, for the reasons that Section 8.2 develops.

Table 6: Questions for evaluating the use of AI by an asset manager and warning signs

Criterion	Question to the manager	Warning sign
Role of AI	Which decisions does the system take alone, and which does a person take?	The manager cannot name the decisions a human remains responsible for
Economic rationale	Why should the signal work, and who is on the other side of the trade?	The explanation refers only to accuracy metrics, not to an economic mechanism
Evidence	Is the track record live or backtested, net of costs, and over which period?	Backtests only; no out-of-sample period; no transaction costs
Multiple testing	How many model variants were tried before this one was chosen?	The question is not understood or the number is not recorded
Data	Could the model have seen the future (look-ahead bias) during training?	General-purpose model used on historical periods without entity masking or cut-off control
Validation	Who validates the model independently, and how often?	The developers validate their own model
Grounding	Are figures in reports generated by the model or taken from verified data?	The manager cannot say
Dependence	Which third-party providers does the system depend on, and what happens if they fail?	A single external model provider without a fallback
Regulation	How are the MiFID II, AI Act and DORA obligations met?	No AI-literacy programme; no inventory of AI systems

Source: Own presentation, drawing on Arnott, Harvey and Markowitz (2019), ESMA (2024) and IOSCO (2021).

8 Challenges: Measurable Benefit, Spurious Correlations, Model Risk and Regulation

8.1 Measurable benefit versus hype

Every general-purpose technology has passed through a period in which investment ran ahead of measurable output. Solow (1987) remarked that the computer age was visible everywhere except in the productivity statistics, and Brynjolfsson, Rock and Syverson (2021) explained the pattern with a productivity J-curve: firms must first accumulate intangible complements, new processes, data, skills and organisational forms, before the technology pays off, so that measured productivity initially falls. Generative AI is at the beginning of this curve. McKinsey & Company (2023) estimated the potential of generative AI at 2.6 to 4.4 trillion US dollars of annual value across industries, with banking among the largest beneficiaries; the survey evidence collected in Figure 4 shows how far practice still is from potential.

The study of KPMG (2026) on the current state of AI in German asset management, based on expert interviews with chief operating officers of asset managers, servicing providers and depositaries and on more than 640 individual assessments, found that every surveyed firm regards AI as strategically important and 75 per cent regard it as critical or decisive for their competitiveness, yet 63 per cent remain in test or proof-of-concept phases

and only 12 per cent have scaled AI successfully with a measurable return on investment. In the words of the study, only a small minority can demonstrate a measurable economic benefit, and the benefit grows with the depth of integration into the value chain rather than with the number of use cases. The Bank of England and the Financial Conduct Authority (2024) found that 75 per cent of UK financial services firms already used AI and a further 10 per cent planned to do so within three years, and Mercer (2024) found that 54 per cent of global investment managers used AI in their investment strategy or asset-class research; in a follow-up survey of 131 managers in early 2026, 55 per cent reported AI integrated in at least one investment process and 27 per cent were at the pilot stage (Mercer, 2026). Adoption, in other words, is nearly universal; measurable benefit is not.[42]

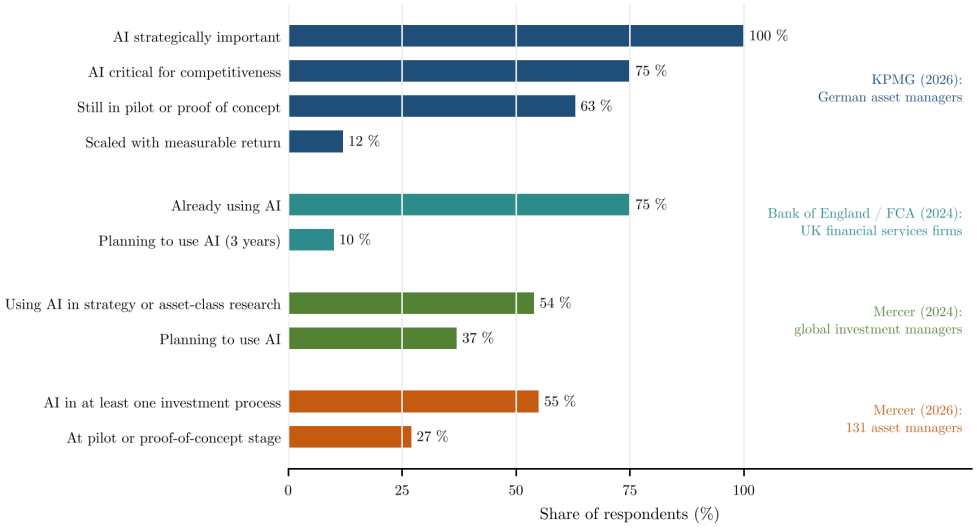


Figure 4: Adoption of AI and measurable benefit in asset management: survey evidence
Source: Own illustration based on KPMG (2026), Bank of England and Financial Conduct Authority (2024), Mercer (2024) and Mercer (2026).

Three reasons explain the gap for investment firms in particular. The first is the small-data problem of Israel, Kelly and Moskowitz (2020): the signal-to-noise ratio in returns is low and non-stationary, so that the spectacular accuracy of language models on linguistic tasks does not carry over to forecasting. The second is crowding: a benefit that every firm obtains from the same provider's model is competed away in prices, as Lo (2004) and Stein (2009) predict. The third is organisational: the measurable benefits lie in operations, reporting and compliance, where they are real but unglamorous, while the promised benefits lie in alpha, where they are glamorous but elusive.[43]

[42] KPMG (2026) attributes the gap to a management approach that counts use cases and usage rather than business outcomes; successful firms, in its account, embed AI in the value chain and manage it against business objectives.
[43] Lo (2004) describes the diffusion of an innovation among competing investors as the mechanism by which excess profits are competed away, in analogy to the competition among species in an ecosystem.

8.2 Spurious correlations and backtest overfitting

The central methodological danger of AI in investing is older than AI. Granger and Newbold (1974) showed that regressions of unrelated trending series produce spurious significance; Lo and MacKinlay (1990) analysed data-snooping biases in tests of asset pricing models; Sullivan, Timmermann and White (1999) and White (2000) showed that the apparent profitability of technical trading rules largely disappears once the number of rules examined is taken into account. Leinweber (2007) made the point memorably by "predicting" the S&P 500 from butter production in Bangladesh. Harvey, Liu and Zhu (2016) catalogued more than 300 published factors and concluded that, given the number of tests, a newly proposed factor should clear a t-statistic of at least three, not two. Flexible models and cheap computation multiply the problem: Bailey et al. (2014) called the practice of selecting the best of many backtests and presenting it as a discovery "pseudo-mathematics and financial charlatanism" and showed that the expected maximum Sharpe ratio of N unrelated strategies grows with the logarithm of N even when every strategy is worthless.

Figure 5 quantifies the argument with the formula of Bailey et al. (2014) for the expected maximum of N independent Sharpe ratios under the null hypothesis that the true Sharpe ratio of every variant is zero. For T years of daily data, the expected maximum annualised in-sample Sharpe ratio is approximately given by equation (4),

$$\mathrm{E}[\max \mathrm{SR}_{\mathrm{N}}] \approx \left[(1-\gamma_{\mathrm{E}})\, \mathrm{Z}^{-1}(1-1/\mathrm{N}) + \gamma_{\mathrm{E}}\, \mathrm{Z}^{-1}(1-1/(\mathrm{N}\,\mathrm{e}))\right] / \sqrt{\mathrm{T}} \qquad (4)$$

where Z^{-1} denotes the inverse of the standard normal distribution function, e Euler's number and $\gamma_{\mathrm{E}} \approx 0.5772$ the Euler–Mascheroni constant; the standard error of the daily Sharpe ratio, $1/\sqrt{(252\ \mathrm{T})}$, and the annualisation factor $\sqrt{252}$ combine to $1/\sqrt{\mathrm{T}}$.[44] Evaluated with equation (4), testing about 45 variants with five years of daily data suffices to obtain an expected best in-sample annualised Sharpe ratio of one; with two years of data, about eight variants suffice, and a thousand variants produce an expected best Sharpe ratio above two (own calculations). A genetic algorithm or a hyperparameter search that evaluates thousands of configurations will therefore always find an impressive backtest, and the impressiveness carries no information about future performance. Bailey and López de Prado (2014) proposed the deflated Sharpe ratio that corrects for the number of trials, and Arnott, Harvey and Markowitz (2019) proposed a research protocol for the era of machine learning whose first requirement is an economic foundation for every hypothesis.

[44] The individual steps of the calculation are documented in the accompanying spreadsheet.

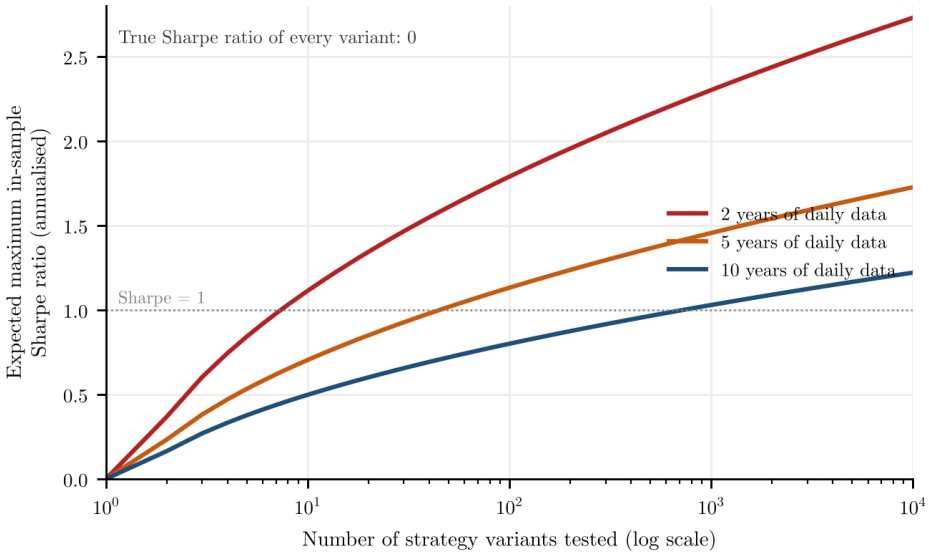


Figure 5: Expected maximum in-sample Sharpe ratio of N worthless strategy variants as a function of N
Source: Own calculations based on Bailey et al. (2014).

Generative AI adds a new form of the same problem. A language model asked to find relationships in a data set will find them, and will produce a plausible economic narrative for each one; the narrative is generated from the pattern rather than the pattern from the narrative. The discipline that Breiman (2001) demanded of the algorithmic culture, honest out-of-sample evaluation, and the discipline that Arnott, Harvey and Markowitz (2019) demanded of the financial culture, an economic rationale stated before the data are examined, are both indispensable, and the second cannot be delegated to the machine that produced the pattern. The brief that motivated this paper put it precisely: pure data models can derive false causalities, which is why an economic foundation and a strict examination of the data are essential.[45]

8.3 Hallucination, look-ahead bias and explainability

Three properties of language models create model risks that classical quantitative models did not have. Hallucination, the generation of fluent statements that are false or unsupported, is intrinsic to models that predict probable text; Ji et al. (2023) survey its forms and causes, and Bender et al. (2021) warned early that the fluency of such models invites readers to ascribe understanding where there is none. Look-ahead bias, discussed in Section 4.3, arises because the training corpus of a foundation model contains the outcomes of the events a researcher wishes to predict; the evidence of Glasserman and Lin (2023) and He et al. (2025) suggests that the bias is modest in practice, but that a model's prior knowledge of well-known firms can distort its judgements. Explainability is weaker than

[45] Breiman (2001) also described the Rashomon effect: many different models fit the same data equally well, so that the choice among them cannot be made by the data alone.

for tree-based or linear models: methods such as the Shapley values of Lundberg and Lee (2017) attribute a prediction to its inputs for structured models, but the attribution of a generated paragraph to the billions of parameters that produced it is not meaningful, and Rudin (2019) argued that high-stakes decisions should rely on inherently interpretable models rather than on post-hoc explanations of black boxes.[46]

Table 7 summarises the risks with their mechanisms and the safeguards that practice and regulation have developed. None of the safeguards is novel; all of them are the model-risk-management principles of the supervisory guidance SR 11-7 of the Federal Reserve and the Office of the Comptroller of the Currency (2011) applied to a new class of model: documented purpose, independent validation, ongoing monitoring, and limits on use outside the validated domain.

Table 7: Risks of AI and generative AI in asset management, their mechanisms and safeguards

Risk	Mechanism	Safeguard	Reference
Backtest overfitting	Selection of the best of many trials; expected maximum Sharpe ratio rises with the number of trials	Pre-registration; deflated Sharpe ratio; holdout periods; record of all trials	Bailey et al. (2014); Harvey, Liu and Zhu (2016)
Spurious correlation	Flexible models fit noise; narratives generated after the fact	Economic rationale stated before testing; out-of-sample validation	Granger and Newbold (1974); Arnott, Harvey and Markowitz (2019)
Look-ahead bias	Training corpus contains outcomes of the test period	Training cut-off before test period; entity masking; point-in-time models	Glasserman and Lin (2023); He et al. (2025)
Hallucination	Fluent but false statements; invented figures and sources	Grounding of all figures in verified data; source citation; human review	Ji et al. (2023)
Opacity	Predictions cannot be attributed to causes	Interpretable models for high-stakes decisions; attribution methods	Rudin (2019); Lundberg and Lee (2017)
Crowding	Many investors use the same models and signals	Diversity of models; capacity limits; monitoring of correlation	Stein (2009); Khandani and Lo (2011)
Third-party dependence	Concentration on few model and cloud providers	Contractual and technical fallbacks; DORA ICT risk management	Financial Stability Board (2024); European Union (2022)
Regime change	Models trained on one regime fail in another	Stress tests; human override; economic priors	Israel, Kelly and Moskowitz (2020)

Source: Own presentation.

8.4 Crowding, homogenisation and systemic risk

The concerns of supervisors go beyond the individual firm. The Financial Stability Board (2017) identified early the potential for new forms of interconnectedness and for the

[46] Shapley values originate in cooperative game theory and distribute a prediction among the input variables according to their average marginal contribution across all possible orderings.

emergence of a few dominant providers; its report of 2024 added the vulnerabilities of third-party dependencies and service-provider concentration, of increased market correlations through similar models, and of cyber and model risk (Financial Stability Board, 2024). The International Monetary Fund (2024) devoted a chapter of its Global Financial Stability Report to AI in capital markets and concluded that AI could make markets more efficient but also faster and more volatile under stress, and that the opacity of models and the reliance on a few providers warrant closer monitoring. The European Central Bank (2024) reached a similar assessment for the euro area, and IOSCO (2021) formulated six measures for market intermediaries and asset managers, from senior-management accountability to the control of outsourced AI.[47]

The academic evidence is still young but points in the same direction. Dou, Goldstein and Ji (2025) showed that reinforcement-learning traders, each maximising its own profit, converge on collusive behaviour that reduces the informativeness of prices, without any communication between them, extending the results of Calvano et al. (2020) on algorithmic pricing to financial markets. Kirilenko et al. (2017) documented how the interaction of automated traders amplified the Flash Crash.[48] If generative AI leads thousands of managers to read the same documents with the same models and reach the same conclusions at the same moment, the efficiency gain of Section 3 and the fragility described here are two faces of the same phenomenon.

8.5 The regulatory framework

The regulatory framework for AI in European asset management has four layers, which Table 8 summarises. The first is the Artificial Intelligence Act, Regulation (EU) 2024/1689, in force since 1 August 2024, which follows a risk-based approach: a short list of prohibited practices, a catalogue of high-risk uses in Annex III, transparency duties for systems that interact with people or generate content, and obligations for providers of general-purpose AI models (European Union, 2024). For asset managers the decisive point is that portfolio management and investment advice are not listed as high-risk uses; the high-risk items relevant to financial institutions are the creditworthiness assessment of natural persons and risk assessment and pricing in life and health insurance. Most investment uses therefore fall under the transparency and AI-literacy duties and under the general obligations of financial regulation rather than under the high-risk regime. The Digital Omnibus on AI, Regulation (EU) 2026/1744, published on 24 July 2026, postponed the application of the Annex III obligations from 2 August 2026 to 2 December 2027 and of the Annex I obligations to 2 August 2028, softened the AI-literacy duty of Article 4 into a duty to take

[47] The measures of IOSCO (2021) cover governance and senior-management accountability, testing and monitoring, staff competence, control of third-party providers, disclosure to clients and regulators, and data quality.

[48] Kirilenko et al. (2017) conclude that high-frequency traders did not trigger the Flash Crash of 6 May 2010 but amplified it by withdrawing liquidity and trading in the direction of the price move.

supporting measures, and introduced relief for small mid-caps (European Union, 2026). Figure 6 shows the resulting timeline.[49]

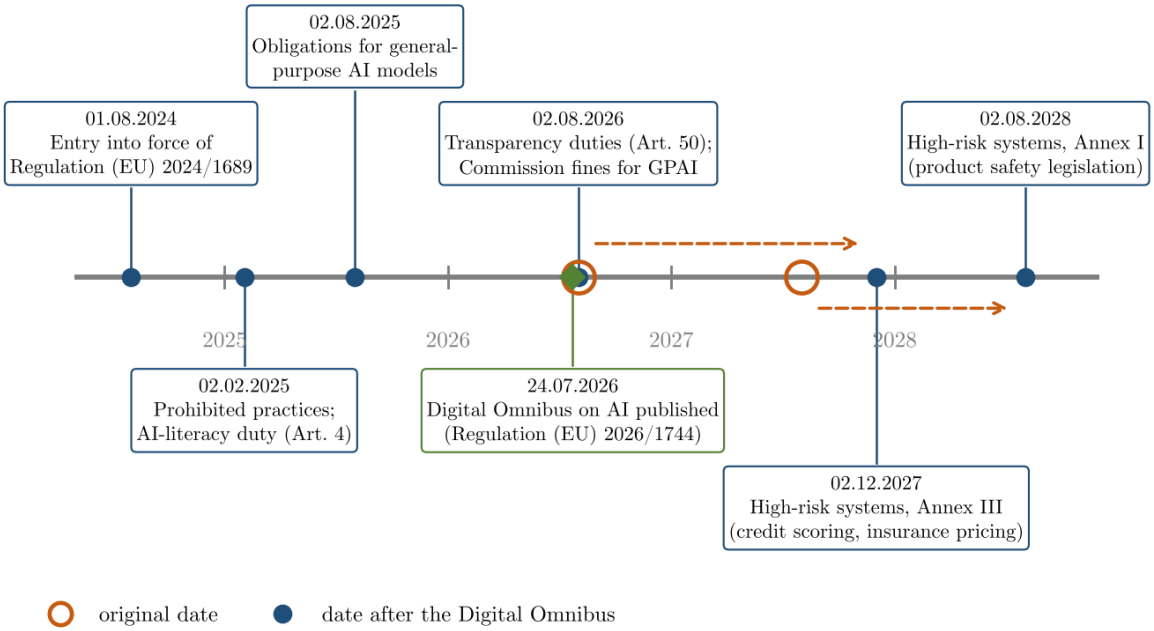


Figure 6: Application dates of the EU Artificial Intelligence Act after the Digital Omnibus of July 2026
Source: Own illustration based on European Union (2024) and European Union (2026).

The second layer is sectoral financial regulation. MiFID II (European Union, 2014) and its delegated acts govern suitability, best execution, conflicts of interest, record keeping and the organisational requirements that apply to any technology; ESMA (2024) has made explicit that they apply to AI. The third layer is operational resilience: the Digital Operational Resilience Act, Regulation (EU) 2022/2554, applicable since 17 January 2025, requires financial entities to manage the risks of ICT third-party providers, which includes the providers of foundation models and of the cloud infrastructure on which they run (European Union, 2022). The fourth layer is data protection: the General Data Protection Regulation restricts the processing of personal data in training and prompting and, in Article 22, the use of solely automated decisions with legal effects on individuals (European Union, 2016). For an Austrian investment firm supervised by the Financial Market Authority, all four layers apply simultaneously, and the AI literacy of staff, the inventory of AI systems in use and the documentation of human oversight are the practical starting points of compliance.[50]

[49] Annex III also lists employment-related uses such as recruitment and performance evaluation, which concern asset managers as employers rather than as investors.
[50] The Austrian Financial Market Authority supervises investment firms under the Securities Supervision Act (WAG 2018), which transposes MiFID II; the obligations described here therefore apply to Austrian firms without further national AI-specific rules.

Table 8: Regulatory framework for AI in European asset management

Instrument	Scope	Relevance for asset managers	Key dates
AI Act, Regulation (EU) 2024/1689, as amended by Regulation (EU) 2026/1744	Risk-based rules for AI systems and general-purpose AI models	Portfolio management not high-risk; transparency and AI-literacy duties; credit scoring and insurance pricing high-risk	In force 1 Aug 2024; GPAI 2 Aug 2025; Annex III 2 Dec 2027; Annex I 2 Aug 2028
MiFID II, Directive 2014/65/EU	Investment services and activities	Suitability, best interest, organisational requirements, record keeping apply to AI-supported services	Applicable since 3 Jan 2018
ESMA public statement on AI (2024)	Retail investment services	Management-body responsibility; testing and monitoring; staff training; client information	30 May 2024
DORA, Regulation (EU) 2022/2554	Digital operational resilience of financial entities	ICT third-party risk management covers model and cloud providers	Applicable since 17 Jan 2025
GDPR, Regulation (EU) 2016/679	Processing of personal data	Training and prompting with client data; Art. 22 automated decisions	Applicable since 25 May 2018
SR 11-7 (Federal Reserve and Office of the Comptroller of the Currency, 2011)	Model risk management (US supervisory guidance)	De facto international standard for validation and governance of models	April 2011

Source: Own presentation based on the cited legal acts and European Union (2026).

8.6 Governance: degrees of autonomy and human oversight

The regulatory layers converge on a single governance question: how much autonomy may an AI system have in a given use, and what form must human oversight take? Figure 7 proposes a simple answer. The degree of autonomy that is acceptable falls as the consequence of an error for the client rises. Summarising documents and drafting texts that a person reviews lie in the zone of assistance, where the cost of an error is low and its detection easy. Monitoring guidelines and generating alerts lie in the middle: the system acts, but a person decides. Autonomous portfolio decisions for a client lie in the zone in which a human decision is mandatory, not because machines are necessarily worse, but because accountability under MiFID II and the fiduciary relationship cannot be transferred to them and because the systemic risks of Section 8.4 grow with the share of decisions taken by similar models.[51]

[51] The AI Act requires human oversight (Art. 14) only for high-risk systems; for other systems the requirement follows from the organisational rules of MiFID II and from civil liability.

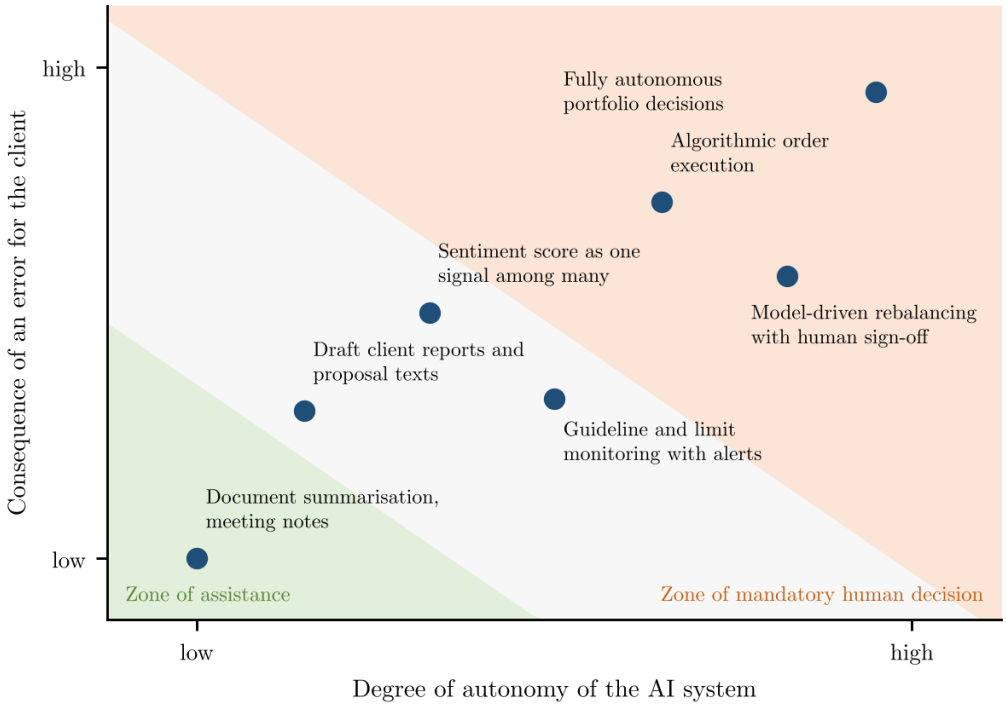


Figure 7: Acceptable degree of autonomy of AI systems in asset management by consequence of an error

Notes: Conceptual framework proposed by the author; the positions of the applications are qualitative assessments, not measurements.

Source: Own illustration.

Operationally, this means three things. Every AI system in use is inventoried with its purpose, its data, its provider and its validated domain. Every output that influences a decision carries its sources and is reviewed by a person whose competence to review it is documented, which is the substance of the AI-literacy duty. And every decision that affects a client is taken, and recorded as taken, by a person. Within these limits, the degree of autonomy can rise with experience; outside them, the experience of August 2007 and of May 2010 suggests that the costs of autonomy are borne by others.

9 Conclusion and Outlook

This paper set out to examine how generative AI, in comparison with traditional quantitative models, changes the information efficiency of equity markets and the practice of asset management. The theoretical argument is straightforward in the framework of Grossman and Stiglitz (1980): classical forecasting models raise the precision of private signals and thereby the incentive to acquire information, whereas generative models lower the cost of processing public information, in particular text, and thereby the frictions that Blankespoor, deHaan and Marinovic (2020) described as disclosure processing costs. The first-order prediction is that generative AI makes blue-chip prices incorporate the qualitative content of earnings calls and filings faster and more completely, shortening the drifts that Bernard and Thomas (1989), Price et al. (2012) and Cohen, Malloy and Nguyen

(2020) documented, while its effect on the long-run informativeness of prices is ambiguous.[52]

The empirical literature supports the premise. From Loughran and McDonald (2011) through Huang, Wang and Yang (2023) to Lopez-Lira and Tang (2023) and Chen, Kelly and Xiu (2022), each generation of textual methods has extracted more return-relevant information from the same documents, and the generative generation does so with unprecedented breadth and speed. The same literature also calls for care: a prominent study was withdrawn by its authors after inconsistencies in the data were found (Kim, Muhn and Nikolaev, 2024), and the tests of Glasserman and Lin (2023) show that a model's prior knowledge of companies can distort its judgements. The research design of Section 5 turns these lessons into a concrete study. We expect, as a hypothesis rather than a result, that the study will show a measurable and declining post-announcement drift associated with LLM-based sentiment surprises in blue-chip stocks, which would constitute direct evidence of rising informational efficiency.

For practice the conclusions are more cautious than the discourse. In fund management, generative AI has become an indispensable reading machine that widens coverage and feeds structured variables into established signal models; it has not produced a generation of superior autonomous funds, and the largest managers use it pervasively while keeping the portfolio decision human. In individual mandates, its benefits lie in suitability documentation, guideline monitoring, tax-aware optimisation and above all in personalised communication, provided every figure is grounded in verified data. The survey of KPMG (2026) is a necessary corrective: all surveyed firms regard AI as strategic, almost two thirds remain in pilots, and roughly one in eight has achieved a measurable return.[53] The statistics of Bailey et al. (2014) are a second corrective: only an economic rationale stated in advance and an honest out-of-sample test separate insight from chance.

The outlook has three dimensions. Technologically, the move from assistants to agents will continue, and open-weight models with documented training corpora will make credible research designs easier. Economically, the advantage of generative AI will migrate from the reading itself, which everyone will do, to the proprietary data that is read and to the judgement that is applied to it; the strongest efficiency gains should appear outside the blue-chip universe, in small caps and in markets with thin coverage, which is where the next empirical work should look. Regulatorily, the European framework has settled on a position that this paper endorses: AI may read, propose and monitor, while a person decides and answers for the decision. Generative AI strengthens economically founded judgement; it does not replace it, and those who forget the difference will discover it, as every previous

[52] Fama (1970) framed efficiency as a statement about the speed and completeness of price adjustment, which is why the drift, not the announcement return, is the natural object of the test.

[53] One in eight corresponds to the 12 per cent of respondents in KPMG (2026) who report scaled AI with a measurable return on investment.

generation of quantitative investors did, in the one period for which their models had no training data.

References

Abis, S. (2020). Man vs. machine: Quantitative and discretionary equity management. Working Paper, SSRN 3717371.

Arnott, R., Harvey, C.R., & Markowitz, H. (2019). A backtesting protocol in the era of machine learning. *Journal of Financial Data Science*, 1(1), 64–74.

Athey, S., & Imbens, G.W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685–725.

Avramov, D., Cheng, S., & Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5), 2587–2619.

Bai, J., Philippon, T., & Savov, A. (2016). Have financial markets become more informative? *Journal of Financial Economics*, 122(3), 625–654.

Bailey, D.H., Borwein, J.M., López de Prado, M., & Zhu, Q.J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5), 458–471.

Bailey, D.H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5), 94–107.

Ball, R., & Brown, P. (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, 6(2), 159–178.

Bank of England and Financial Conduct Authority (2024). Artificial intelligence in UK financial services – 2024. London, 21 November 2024.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49(3), 307–343.

Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bernard, V.L., & Thomas, J.K. (1989). Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27 (Supplement), 1–36.

Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *Review of Financial Studies*, 34(2), 1046–1089.

Black, F., & Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, 48(5), 28–43.

Blankespoor, E., deHaan, E., & Marinovic, I. (2020). Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics*, 70(2–3), 101344.

Bloomberg (2024). Bridgewater launches $2 billion fund run by machine learning. Bloomberg News, 1 July 2024.

Boehmer, E., Musumeci, J., & Poulsen, A.B. (1991). Event-study methodology under conditions of event-induced variance. *Journal of Financial Economics*, 30(2), 253–272.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231.

Brown, S.J., & Warner, J.B. (1985). Using daily stock returns: The case of event studies. *Journal of Financial Economics*, 14(1), 3–31.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2), 889–942.

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1), 333–372.

Bybee, L., Kelly, B., Manela, A., & Xiu, D. (2024). Business news and business cycles. *Journal of Finance*, 79(5), 3105–3147.

Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.

Cao, S., Jiang, W., Yang, B., & Zhang, A.L. (2023). How to talk when a machine is listening: Corporate disclosure in the age of AI. *Review of Financial Studies*, 36(9), 3603–3642.

Carhart, M.M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1), 57–82.

Chen, Y., Kelly, B.T., & Xiu, D. (2022). Expected returns and large language models. Working Paper, Yale University and University of Chicago, SSRN 4416687.

CNBC (2023). JPMorgan is developing a ChatGPT-like A.I. service that gives investment advice. CNBC, 25 May 2023.

CNBC (2025). Goldman Sachs rolls out an AI assistant for its employees as artificial intelligence sweeps Wall Street (H. Son). CNBC, 21 January 2025.

Cochrane, J.H. (2011). Presidential address: Discount rates. *Journal of Finance*, 66(4), 1047–1108.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. *Journal of Finance*, 75(3), 1371–1415.

Cong, L.W., Tang, K., Wang, J., & Zhang, Y. (2021). AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI. Working Paper, Cornell University, SSRN 3554486.

Correlation One (2026). The analyst knowledge locked inside company calls is your firm's most underutilized asset. Correlation One blog, 18 May 2026.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance*, 66(5), 1461–1499.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. *Journal of Finance*, 53(6), 1839–1885.

Dell'Acqua, F., McFowland, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., et al. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard Business School Working Paper 24-013, SSRN 4573321.

Dessaint, O., Foucault, T., & Frésard, L. (2024). Does alternative data improve financial forecasting? The horizon effect. *Journal of Finance*, 79(3), 2237–2287.

Deutsche Börse (2021). DAX mit zehn neuen Mitgliedern [DAX with ten new members]. Press release, Zug, 3 September 2021.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

Dou, W.W., Goldstein, I., & Ji, Y. (2025). AI-powered trading, algorithmic collusion, and price efficiency. NBER Working Paper 34054.

Druz, M., Petzev, I., Wagner, A.F., & Zeckhauser, R.J. (2020). When managers change their tone, analysts and investors change their tune. *Financial Analysts Journal*, 76(2), 47–69.

Dugast, J., & Foucault, T. (2018). Data abundance and asset price informativeness. *Journal of Financial Economics*, 130(2), 367–391.

Eisfeldt, A.L., Schubert, G., & Zhang, M.B. (2023). Generative AI and firm values. NBER Working Paper 31222.

ESMA (2024). Public statement on the use of artificial intelligence (AI) in the provision of retail investment services. European Securities and Markets Authority, ESMA35-335435667-5924, 30 May 2024.

European Central Bank (2024). The rise of artificial intelligence: Benefits and risks for financial stability. *Financial Stability Review*, May 2024, special feature (15 May 2024).

European Union (2014). Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments (MiFID II). *Official Journal of the European Union*, L 173, 349–496.

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). *Official Journal of the European Union*, L 119, 1–88.

European Union (2017). Commission Delegated Regulation (EU) 2017/565 of 25 April 2016 supplementing Directive 2014/65/EU as regards organisational requirements and operating conditions for investment firms. *Official Journal of the European Union*, L 87, 31 March 2017, 1–83.

European Union (2022). Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (DORA). *Official Journal of the European Union*, L 333, 1–79.

European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 12 July 2024.

European Union (2026). Regulation (EU) 2026/1744 of the European Parliament and of the Council amending Regulation (EU) 2024/1689 (Digital Omnibus on AI). *Official Journal of the European Union*, 24 July 2026.

Fama, E.F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2), 383–417.

Fama, E.F. (1991). Efficient capital markets: II. *Journal of Finance*, 46(5), 1575–1617.

Fama, E.F., & French, K.R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.

Farboodi, M., & Veldkamp, L. (2020). Long-run growth of financial data technology. *American Economic Review*, 110(8), 2485–2523.

Federal Reserve and Office of the Comptroller of the Currency (2011). Supervisory guidance on model risk management. SR Letter 11-7, 4 April 2011.

Fedyk, A., & Hodson, J. (2023). When can the market identify old news? *Journal of Financial Economics*, 149(1), 92–113.

Financial Stability Board (2017). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. Basel, 1 November 2017.

Financial Stability Board (2024). The financial stability implications of artificial intelligence. Basel, 14 November 2024.

Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science*, 68(7), 5514–5532.

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, 33(5), 2326–2377.

García, D. (2013). Sentiment during recessions. *Journal of Finance*, 68(3), 1267–1300.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574.

Glasserman, P., & Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. Working Paper, Columbia University, arXiv:2309.17322.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Granger, C.W.J., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2(2), 111–120.

Grossman, S.J., & Stiglitz, J.E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3), 393–408.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), 2223–2273.

Hansen, A.L., & Kazinnik, S. (2023). Can ChatGPT decipher Fedspeak? Working Paper, Federal Reserve Bank of Richmond, SSRN 4399406.

Harvey, C.R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

He, S., Lv, L., Manela, A., & Wu, J. (2025). Chronologically consistent large language models. Working Paper, arXiv:2502.21206.

Heston, S.L., & Sinha, N.R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, 73(3), 67–83.

Hirshleifer, D., & Teoh, S.H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1–3), 337–386.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press.

Huang, A.H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2), 806–841.

International Monetary Fund (2024). Advances in artificial intelligence: Implications for capital market activities. *Global Financial Stability Report*, October 2024, Chapter 3.

IOSCO (2021). The use of artificial intelligence and machine learning by market intermediaries and asset managers. Final Report FR06/2021, International Organization of Securities Commissions, September 2021.

Israel, R., Kelly, B.T., & Moskowitz, T.J. (2020). Can machines "learn" finance? *Journal of Investment Management*, 18(2), second quarter 2020.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3), 712–729.

Jha, M., Qian, J., Weber, M., & Yang, B. (2024). ChatGPT and corporate policies. NBER Working Paper 32161.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.

JPMorgan Chase (2024). Annual Report 2023: Chairman and CEO letter to shareholders. JPMorgan Chase & Co., April 2024.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Ke, Z.T., Kelly, B.T., & Xiu, D. (2019). Predicting returns with text data. NBER Working Paper 26186.

Kelly, B., Malamud, S., & Zhou, K. (2024). The virtue of complexity in return prediction. *Journal of Finance*, 79(1), 459–503.

Kelly, B., & Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*, 13(3–4), 205–363.

Khandani, A.E., & Lo, A.W. (2011). What happened to the quants in August 2007? Evidence from factors and transactions data. *Journal of Financial Markets*, 14(1), 1–46.

Kim, A., Muhn, M., & Nikolaev, V. (2023). Bloated disclosures: Can ChatGPT help investors process information? Chicago Booth Research Paper 23-07, SSRN 4425527.

Kim, A., Muhn, M., & Nikolaev, V. (2024). Financial statement analysis with large language models. Working Paper, arXiv:2407.17866 (withdrawn by the authors on 20 February 2025).

Kirilenko, A., Kyle, A.S., Samadi, M., & Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. *Journal of Finance*, 72(3), 967–998.

Kolari, J.W., & Pynnönen, S. (2010). Event study testing with cross-sectional correlation of abnormal returns. *Review of Financial Studies*, 23(11), 3996–4025.

Kothari, S.P., & Warner, J.B. (2007). Econometrics of event studies. In B.E. Eckbo (Ed.), *Handbook of Empirical Corporate Finance* (Vol. 1, pp. 3–36). Elsevier.

KPMG (2026). KI im Asset Management: Zwischen strategischer Priorität und echter Wirkung. Studie „Current State of AI im Asset Management“. KPMG AG Wirtschaftsprüfungsgesellschaft, Frankfurt am Main (accessed 29 October 2026).

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.

Kyle, A.S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335.

Larcker, D.F., & Zakolyukina, A.A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2), 495–540.

Leinweber, D.J. (2007). Stupid data miner tricks: Overfitting the S&P 500. *Journal of Investing*, 16(1), 15–22.

Lo, A.W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management*, 30(5), 15–29.

Lo, A.W., & MacKinlay, A.C. (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, 3(3), 431–467.

López de Prado, M. (2016). Building diversified portfolios that outperform out of sample. *Journal of Portfolio Management*, 42(4), 59–69.

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. Working Paper, University of Florida, arXiv:2304.07619 (revised version October 2025).

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187–1230.

Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

MacKinlay, A.C. (1997). Event studies in economics and finance. *Journal of Economic Literature*, 35(1), 13–39.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77–91.

Mayew, W.J., & Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance. *Journal of Finance*, 67(1), 1–43.

McCarthy, J., Minsky, M.L., Rochester, N., & Shannon, C.E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Reprinted in *AI Magazine*, 27(4), 2006, 12–14.

McKinsey & Company (2023). The economic potential of generative AI: The next productivity frontier (M. Chui, E. Hazan, R. Roberts, A. Singla, K. Smaje, A. Sukharevsky, L. Yee and R. Zemmel). McKinsey & Company, 14 June 2023.

Mercer (2024). AI integration in investment management: 2024 global manager survey. Mercer Investments, March 2024.

Mercer (2026). Moving beyond the AI pitch: Asset managers' use of AI. 2026 AI in Asset Management Survey. Mercer Investments.

Misonzhnik, E. (2025). Inside BlackRock's approach to systematic investing. *WealthManagement.com*, 27 October 2025.

Morgan Stanley (2024). Morgan Stanley Wealth Management announces latest game-changing addition to suite of GenAI tools. Press release, 26 June 2024.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192.

OpenAI (2026). How Balyasny Asset Management built an AI research engine. Customer story, OpenAI, 6 March 2026.

Petersen, M.A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies*, 22(1), 435–480.

Price, S.M., Doran, J.S., Peterson, D.R., & Bliss, B.A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4), 992–1011.

private banking magazin (2017). Für Privatanleger: Acatis legt KI-gesteuerten Fonds auf. private banking magazin, 5 July 2017.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.

Rupprechter, R. (2026). Man or machine? Where artificial intelligence adds value in asset management and where human judgement prevails: Asset allocation and security selection by naive and professional investors. Working Paper 33, R&B Wealth Management, Dornbirn, September 2026.

Securities and Exchange Commission (2000). Selective disclosure and insider trading. Final rule, Release No. 33-7881, 15 August 2000.

Sharpe, W.F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3), 425–442.

Solow, R.M. (1987). We'd better watch out. *New York Times Book Review*, 12 July 1987, 36.

Stein, J.C. (2009). Presidential address: Sophisticated investors and market efficiency. *Journal of Finance*, 64(4), 1517–1548.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54(5), 1647–1691.

Tetlock, P.C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168.

Thieringer, J. (2025). Wie Acatis „kleine Warren Buffetts für jeden Sektor" baut. Interview with Kevin Endler. DAS INVESTMENT, 16 May 2025.

ThinkAdvisor (2017). AI-powered equity ETF launches on NYSE Arca: Portfolio products. ThinkAdvisor, 23 October 2017.

Union Investment (2024). Union Investment startet Vermögensverwaltung mit Künstlicher Intelligenz. Press release, Frankfurt am Main, 2024.

Varian, H.R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

White, H. (1988). Economic prediction using neural networks: The case of IBM daily stock returns. *Proceedings of the IEEE International Conference on Neural Networks*, 2, 451–458.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. Working Paper, Bloomberg, arXiv:2303.17564.