Künstliche Intelligenz und generative KI im Asset Management

Große Sprachmodelle, autonome Agenten und die Informationseffizienz der Aktienmärkte

Roland Rupprechter[1]

Oktober 2026

Zusammenfassung

Diese Arbeit untersucht, wie generative künstliche Intelligenz, insbesondere große Sprachmodelle (LLMs) und autonome KI-Agenten, die Informationseffizienz der Aktienmärkte und die Praxis des Asset Managements im Vergleich zu klassischen quantitativen Modellen verändert. Ausgehend von Fama (1970) und Grossman und Stiglitz (1980) argumentieren wir, dass generative KI die Kosten der Verarbeitung unstrukturierter Information senkt, nicht die Kosten der Beschaffung privater Information; ihre Wirkung erster Ordnung ist eine schnellere Einpreisung öffentlicher Texte und eine kürzere Kursdrift nach Ankündigungen.

Wir sichten die empirische Literatur von den Wörterbuchmethoden von Loughran und McDonald (2011) bis zu den LLM-basierten Studien zur Renditeprognose von Lopez-Lira und Tang (2023) und Chen, Kelly und Xiu (2022) und entwickeln ein vollständiges Forschungsdesign für eine Ereignisstudie zu KI-gestützten Sentiment-Maßen und Kursreaktionen auf Earnings Calls von Blue-Chip-Aktien.

Die Evidenz stützt ein nüchternes Fazit: Ein messbarer wirtschaftlicher Nutzen ist bislang auf eine Minderheit der Häuser beschränkt, Scheinkorrelationen und Backtest-Überanpassung bleiben die zentrale methodische Gefahr, und die EU-KI-Verordnung bildet zusammen mit MiFID II einen Governance-Rahmen, innerhalb

[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. *Erklärung zum Einsatz von KI-Tools:* Diese Arbeit entstand unter punktuellem Einsatz der KI Claude (Anthropic). Inhaltliche Kernleistungen – darunter die Forschungsfrage, das Ziel, die Methodik, die Datenanalyse, sämtliche Berechnungen sowie das Fachwissen und das Layout – wurden vom Autor komplett eigenständig erbracht. Claude diente unter Anleitung als Assistenzwerkzeug zur Unterstützung bei der Recherche, zur Textkürzung, zur Überprüfung von Herleitungen sowie zum Abgleich von Quellen. Manche KI-generierte Textpassagen wurden vom Autor intensiv geprüft und erst nach Erreichen der inhaltlichen Ziele integriert. Zuletzt wurde das System für das Lektorat (Rechtschreibung und akademischer Stil) genutzt. Der Autor trägt die alleinige Verantwortung für den Inhalt der Arbeit.

dessen der menschliche Portfoliomanager die letzte Entscheidung behält. Generative KI stärkt ökonomisch fundiertes Anlageurteil, ersetzt es aber nicht.

Schlagwörter: Künstliche Intelligenz · Generative KI · Große Sprachmodelle · Markteffizienz · Sentiment-Analyse · Ereignisstudie · Earnings Calls · Asset Management · Fondsmanagement · Backtest-Überanpassung · EU-KI-Verordnung.
JEL-Codes: C45, C58, G11, G12, G14, G17, G23, O33.

Nicht-technische Zusammenfassung

Künstliche Intelligenz hat die Investmentbranche in zwei Wellen erreicht. Die erste Welle, die in den 1980er-Jahren begann, bestand aus statistischen Modellen und Verfahren des maschinellen Lernens, die in Kursen, Bilanzen und makroökonomischen Daten nach Mustern suchen und daraus Prognosen für Renditen und Risiken ableiten. Die zweite Welle, die mit der Veröffentlichung von ChatGPT im November 2022 begann, besteht aus generativen Modellen, die Texte lesen, zusammenfassen und schreiben und die als sogenannte Agenten zunehmend selbstständig handeln können. Diese Arbeit fragt, was die zweite Welle für die Aktienmärkte und für jene Menschen verändert, die fremdes Vermögen verwalten.

Unsere zentrale Frage lautet, ob generative KI die Aktienmärkte effizienter macht, ob also Kurse neue Informationen schneller und vollständiger aufnehmen. Die Antwort, die sich aus der wissenschaftlichen Literatur und aus den ersten Jahren der Praxis ergibt, ist ein eingeschränktes Ja. Große Sprachmodelle senken die Kosten des Lesens langer und komplexer Dokumente wie Geschäftsberichte und Mitschriften von Earnings Calls, und die Märkte reagieren schneller auf Informationen, die zwar immer öffentlich, aber teuer zu verarbeiten waren. Zugleich werden die Signale vieler Anleger einander ähnlicher, wenn sie dieselben Modelle verwenden; das Gedränge in denselben Positionen nimmt zu, und der Vorteil des einzelnen Nutzers schrumpft.

Um die Frage greifbar zu machen, beschreiben wir Schritt für Schritt, wie ein Forscher die Wirkung KI-gestützter Sentiment-Analyse auf die Aktienkurse rund um Earnings Calls großer Unternehmen prüfen würde: welche Unternehmen zu untersuchen sind, wie sich der Tonfall eines Calls mit einem Sprachmodell messen lässt, wie abnormale Renditen berechnet werden und welche Fehler, vor allem der Einsatz von Modellen, die die Zukunft bereits gesehen haben, vermieden werden müssen. Wir beschreiben zudem in einfachen Worten, wo KI heute im Fondsmanagement und in der Verwaltung individueller Kundenmandate eingesetzt wird, vom Research und der Portfoliokonstruktion bis zum Reporting und zur Compliance.

Für Anleger ist das praktische Fazit dreifach. Erstens ist KI eine mächtige Hilfe bei der Verarbeitung von Information, doch eine Befragung deutscher Asset Manager zeigt, dass bisher nur eine kleine Minderheit einen messbaren wirtschaftlichen Ertrag daraus erzielt hat. Zweitens gilt: Je flexibler ein Modell, desto leichter wird Zufall mit Erkenntnis verwechselt; wir zeigen, dass das Testen genügend vieler Strategievarianten immer einen eindrucksvollen Backtest hervorbringt, selbst wenn keine einzige Variante etwas taugt. Drittens verlangen die europäische Regulierung und die treuhänderische Pflicht des Vermögensverwalters, dass ein Mensch für jede Anlageentscheidung verantwortlich bleibt. Wer KI als disziplinierten Assistenten einsetzt und nicht als Orakel, wird am meisten profitieren.

Inhaltsverzeichnis

Nicht-technische Zusammenfassung 3
1 Einleitung 6
2 Begriffliche Grundlagen: von klassischen quantitativen Modellen zur generativen KI 8
2.1 Künstliche Intelligenz, maschinelles Lernen und große Sprachmodelle 8
2.2 Klassische KI: Mustererkennung und Prognose 9
2.3 Generative KI und autonome Agenten 10
2.4 Meilensteine der KI im Finanzwesen 12
3 Informationseffizienz der Aktienmärkte im Zeitalter der KI 14
3.1 Die Effizienzmarkthypothese und das Grossman-Stiglitz-Paradoxon 14
3.2 Verarbeitungskosten, begrenzte Aufmerksamkeit und textliche Information 15
3.3 Datenfülle und Informationsgehalt der Kurse 16
3.4 Hypothesen zur Wirkung generativer KI 17
4 Empirische Evidenz: Text, Sentiment und Aktienrenditen 18
4.1 Von Wörterbüchern zu Transformern 18
4.2 Earnings Calls als Informationsereignis 19
4.3 Evidenz aus großen Sprachmodellen 20
5 Forschungsdesign: KI-gestützte Sentiment-Analyse und Kursreaktionen auf Earnings Calls von Blue-Chip-Aktien 23
5.1 Forschungsfrage und Hypothesen 23
5.2 Stichprobe und Daten 23
5.3 Messung des Sentiments 24
5.4 Methodik der Ereignisstudie 25
5.5 Erwartete Ergebnisse, Fallstricke und Robustheit 28
6 Anwendungen I: Fondsmanagement 28
6.1 Signalgewinnung und Alpha-Research 28
6.2 Portfoliokonstruktion und Risikomanagement 29
6.3 Research und Datenfilterung mit Natural Language Processing 30
6.4 Handel, Betrieb und Kundenkommunikation 31
6.5 Praktische Beispiele 31
7 Anwendungen II: Asset Management für Einzelmandate 33
7.1 Anlagerichtlinien, Eignung und Überwachung von Beschränkungen 33
7.2 Personalisiertes Reporting und Kundenkommunikation 34
7.3 Steuerbewusste Optimierung und Rebalancing 35
7.4 KI-Assistenten für Berater und die treuhänderische Rolle des Menschen 35
7.5 Fragen, die Kunden von Einzelmandaten stellen sollten 37

8 Herausforderungen: messbarer Nutzen, Scheinkorrelationen, Modellrisiko und Regulierung 38
8.1 Messbarer Nutzen versus Hype 38
8.2 Scheinkorrelationen und Backtest-Überanpassung 40
8.3 Halluzination, Look-ahead-Bias und Erklärbarkeit 42
8.4 Crowding, Homogenisierung und Systemrisiko 43
8.5 Der regulatorische Rahmen 44
8.6 Governance: Autonomiegrade und menschliche Aufsicht 46
9 Fazit und Ausblick 47
Literaturverzeichnis 50

1 Einleitung

Wenige Technologien haben das Asset Management mit so hohen Erwartungen erreicht wie die generative künstliche Intelligenz. Innerhalb von drei Jahren nach der öffentlichen Freigabe von ChatGPT im November 2022 wurden große Sprachmodelle (LLMs) in die Research-Plattformen der größten systematischen Manager, in die Arbeitsumgebung der Berater globaler Vermögensverwalter und in die Compliance-Funktionen der Banken eingebettet, während die ersten weitgehend von maschinellem Lernen geführten Fonds Leistungsnachweise über mehrere Jahre vorweisen können.[2] Die Begeisterung beruht auf einem alten Versprechen. Seit Markowitz (1952) den Austausch zwischen erwarteter Rendite und Risiko formalisiert hat und seit Sharpe (1964) und Fama (1970) beschrieben haben, wie wettbewerbliche Märkte Information in Kurse einarbeiten, sucht die Investmentbranche nach systematischen Wegen, mehr Information schneller und mit weniger Verhaltensfehlern zu verarbeiten, als das menschliche Urteil erlaubt. Generative KI scheint genau das zu liefern: einen Leser, der nie ermüdet, der zehntausend Seiten Berichte in Minuten verarbeitet und in jeder Sprache eine schlüssige Zusammenfassung schreibt.

Doch dasselbe Versprechen begleitete jede frühere Welle quantitativer Innovation, von den Expertensystemen der 1980er-Jahre über die Experimente mit neuronalen Netzen von White (1988) und den von Harvey, Liu und Zhu (2016) dokumentierten Faktorzoo[3] bis zu den Deep-Learning-Modellen der Kapitalmarkttheorie von Gu, Kelly und Xiu (2020). Jede Welle brachte echte Fortschritte, und jede brachte Scheinentdeckungen, überfüllte Positionen und gelegentlich Krisen hervor, etwa die von Khandani und Lo (2011) analysierte Quant-Krise vom August 2007. Die Frage dieser Arbeit lautet daher nicht, ob generative KI von Bedeutung ist, sondern worin sie sich von den klassischen quantitativen Modellen unterscheidet, die ihr vorausgingen, und was dieser Unterschied für die Informationseffizienz der Aktienmärkte und für die Praxis der Verwaltung von Publikumsfonds und Einzelmandaten bedeutet.

Im Mittelpunkt steht der Gegensatz zweier Modellfamilien. Klassische KI im Sinne des statistischen Lernens ist für Mustererkennung und Prognose gebaut: Aus strukturierten Eingaben schätzt sie einen bedingten Erwartungswert von Renditen, Volatilitäten oder Ausfallwahrscheinlichkeiten und übergibt die Schätzung an einen Optimierer. Generative KI im Sinne transformerbasierter Sprachmodelle ist für die Erzeugung von Inhalten gebaut: Sie liest und schreibt, sie fasst zusammen und schließt in natürlicher Sprache, und wenn sie Daten und andere Programme abfragen kann, plant und erledigt sie als Agent

[2] Der Begriff „künstliche Intelligenz“ (KI) wird in dieser Arbeit im Sinne von Art. 3 Abs. 1 der EU-KI-Verordnung verwendet: ein maschinengestütztes System, das aus seinen Eingaben ableitet, wie es Ausgaben wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erzeugt.

[3] Der Ausdruck „Faktorzoo“ geht auf die Presidential Address von Cochrane (2011) zurück, der fragte, welche der Hunderten veröffentlichten Renditeprädiktoren eigenständige Informationsquellen sind.

mehrstufige Aufgaben. Die ökonomischen Folgen dieser beiden Fähigkeiten sind nicht dieselben. Ein besseres Prognosemodell erhöht die Präzision privater Signale; eine bessere Lesemaschine senkt die Kosten der Verarbeitung öffentlicher Information. Im Rahmen von Grossman und Stiglitz (1980) erhöht Ersteres den Anreiz, Information zu beschaffen, während Letzteres jene Reibungen verringert, die öffentliche Information daran hindern, sich vollständig in den Kursen niederzuschlagen, Reibungen, die Hirshleifer und Teoh (2003) sowie Blankespoor, deHaan und Marinovic (2020) als begrenzte Aufmerksamkeit und Verarbeitungskosten von Offenlegungen beschrieben haben.

Die Literatur zur Textanalyse im Finanzwesen liefert die empirische Grundlage. Tetlock (2007) zeigte, dass der Tonfall einer Zeitungskolumne kurzfristige Marktbewegungen vorhersagt; Loughran und McDonald (2011) wiesen nach, dass allgemeine Wortlisten Finanzsprache falsch klassifizieren, und entwickelten ein finanzspezifisches Wörterbuch; Gentzkow, Kelly und Taddy (2019) sowie Loughran und McDonald (2016) sichteten das Feld auf seinem Weg vom Wörterzählen zum überwachten Lernen. Die Transformer-Architektur von Vaswani et al. (2017) und die vortrainierten Modelle von Devlin et al. (2019) und Brown et al. (2020) machten es anschließend möglich, die Bedeutung von Finanztexten zu messen und nicht nur ihr Vokabular. Huang, Wang und Yang (2023) zeigten, dass ein auf Finanztexte abgestimmtes Modell (FinBERT) Wörterbücher und klassisches maschinelles Lernen bei der Klassifikation des Sentiments von Sätzen aus Analystenberichten übertrifft; Lopez-Lira und Tang (2023), Chen, Kelly und Xiu (2022) sowie Jha et al. (2024) zeigten, dass allgemeine LLMs renditerelevante Information aus Schlagzeilen, Nachrichtenartikeln und Transkripten von Earnings Calls gewinnen. Zugleich könnten Modelle, die mit Textkorpora aus dem Internet trainiert wurden, die Zukunft gewissermaßen schon gelesen haben, was die Frage nach einem Look-ahead-Bias aufwirft; die Tests von Glasserman und Lin (2023) sowie He et al. (2025) deuten darauf hin, dass dieser Bias geringer ist als befürchtet, während die Ablenkung durch das Vorwissen eines Modells über bekannte Unternehmen größer sein kann.

Diese Arbeit leistet vier Beiträge. Erstens liefert sie einen begrifflichen Vergleich klassischer und generativer KI entlang der Stufen der Wertschöpfungskette des Asset Managements und leitet prüfbare Hypothesen zu ihrer Wirkung auf die Markteffizienz ab. Zweitens sichtet sie die empirische Evidenz zu Text, Sentiment und Renditen mit besonderem Augenmerk auf Earnings Calls, das informationsreichste wiederkehrende Ereignis im Leben eines börsennotierten Unternehmens. Drittens entwickelt sie ein vollständiges Forschungsdesign für die Beispielfrage, die diese Studie veranlasst hat, nämlich den Einfluss KI-gestützter Sentiment-Analyse auf Kursreaktionen bei Earnings Calls von Blue-Chip-Aktien, einschließlich Stichprobenbildung, Sentiment-Messung, Ökonometrie der Ereignisstudie und jener Fallstricke, die ein belastbares von einem scheinbaren Ergebnis unterscheiden. Viertens dokumentiert sie praktische Anwendungen für Fonds und für Einzelmandate und konfrontiert sie mit der ernüchternden Befragungsevidenz zum messbaren Nutzen, mit der Statistik der Backtest-Überanpassung

und mit dem europäischen Regulierungsrahmen in der Fassung des Digital Omnibus vom Juli 2026. Die Hypothesen in Kapitel 3.4, das Forschungsdesign in Kapitel 5, der Governance-Rahmen in Kapitel 8.6 und die Prüffragen in Tabelle 6 sind eigene Vorschläge des Verfassers; sie sind keine empirischen Ergebnisse und im Text als solche gekennzeichnet.

Die Arbeit ist wie folgt gegliedert. Kapitel 1 (das vorliegende Kapitel) dient als Einleitung, umreißt die Forschungsfrage und sichtet die bisherige Literatur. Kapitel 2 legt die begrifflichen Grundlagen und stellt klassische quantitative Modelle generativer KI und autonomen Agenten gegenüber. Kapitel 3 analysiert die Informationseffizienz der Aktienmärkte im Zeitalter der KI und formuliert Hypothesen. Kapitel 4 sichtet die empirische Evidenz zu Text, Sentiment und Renditen. Kapitel 5 entwickelt das Forschungsdesign für eine Ereignisstudie zu KI-gestütztem Sentiment und Reaktionen auf Earnings Calls. Die Kapitel 6 und 7 untersuchen Anwendungen im Fondsmanagement und in der Verwaltung von Einzelmandaten. Kapitel 8 erörtert die Herausforderungen des messbaren Nutzens, der Scheinkorrelationen, des Modellrisikos und der Regulierung, und Kapitel 9 enthält das Fazit und einen Ausblick.

2 Begriffliche Grundlagen: von klassischen quantitativen Modellen zur generativen KI

2.1 Künstliche Intelligenz, maschinelles Lernen und große Sprachmodelle

Das Vokabular des Fachgebiets ist geschichtet. Künstliche Intelligenz ist der weiteste Begriff und bezeichnet jedes System, das Aufgaben erledigt, die bei menschlicher Ausführung Intelligenz erfordern würden. Maschinelles Lernen ist jene Teilmenge der KI, bei der die Regeln nicht programmiert, sondern aus Daten geschätzt werden; Hastie, Tibshirani und Friedman (2009) beschreiben seinen statistischen Kern als Schätzung einer Funktion, die Eingaben so mit Ausgaben verknüpft, dass der erwartete Prognosefehler minimiert wird, und Breiman (2001) stellte diese „algorithmische Modellierungskultur“ der Datenmodellierungskultur der klassischen Ökonometrie gegenüber. Deep Learning ist nach Goodfellow, Bengio und Courville (2016) maschinelles Lernen mit vielschichtigen neuronalen Netzen, die ihre eigenen Repräsentationen der Daten lernen, statt auf von Hand gebildete Merkmale angewiesen zu sein.[4]

Große Sprachmodelle sind tiefe neuronale Netze vom Transformer-Typ, den Vaswani et al. (2017) eingeführt haben. Sie werden mit sehr großen Textkorpora darauf trainiert,

[4] Die Unterscheidung zwischen von Hand gebildeten Merkmalen und gelernten Repräsentationen ist für den Vergleich mit klassischen Faktormodellen zentral, deren Charakteristika (Value, Momentum, Qualität, Größe) ökonomische Konstrukte sind, die der Forscher wählt.

das jeweils nächste Wortstück vorherzusagen, und erwerben dabei eine statistische Repräsentation von Syntax, Semantik und einem beträchtlichen Umfang an Weltwissen. Devlin et al. (2019) zeigten mit BERT, dass sich solche vortrainierten Repräsentationen mit wenigen gekennzeichneten Daten an spezifische Aufgaben anpassen lassen; Brown et al. (2020) zeigten mit GPT-3, dass hinreichend große Modelle neue Aufgaben aus einer Handvoll Beispiele in der Eingabe erledigen. Im Finanzwesen trainierten Wu et al. (2023) BloombergGPT, ein Modell mit 50 Milliarden Parametern, mit einem Korpus aus Finanzdokumenten und allgemeinen Texten, und Huang, Wang und Yang (2023) veröffentlichten FinBERT, ein auf Finanzkommunikation abgestimmtes Modell. Der Begriff „generative KI“ bezeichnet den Einsatz solcher Modelle zur Erzeugung neuer Inhalte, von Text, Programmcode, Bildern oder strukturierten Daten, im Gegensatz zum diskriminativen Einsatz, bei dem bestehende Inhalte klassifiziert oder bewertet werden.

2.2 Klassische KI: Mustererkennung und Prognose

Klassische KI im Asset Management ist im Kern angewandte Statistik mit flexiblen Funktionsformen. Ihre Eingaben sind strukturiert: Kurse und Umsätze, Rechnungslegungsgrößen, Analystenschätzungen, makroökonomische Reihen und zunehmend alternative Daten wie Kreditkartentransaktionen, Satellitenbilder oder Stellenanzeigen. Ihre Ausgaben sind Prognosen, für erwartete Renditen, Varianzen, Korrelationen, Ausfallwahrscheinlichkeiten oder die Wahrscheinlichkeit eines Regimewechsels. Mullainathan und Spiess (2017) sahen den komparativen Vorteil des maschinellen Lernens in der Lösung von Prognoseproblemen, im Gegensatz zur Schätzung kausaler Parameter, und Varian (2014) sowie Athey und Imbens (2019) katalogisierten die Verfahren, die aus der Informatik in die Ökonomie gewandert sind: penalisierte Regression, Random Forests (Breiman, 2001), Gradient Boosting und neuronale Netze.[5]

Den entscheidenden empirischen Nachweis für Aktien erbrachten Gu, Kelly und Xiu (2020), die diese Verfahren bei der Prognose des Querschnitts monatlicher US-Aktienrenditen aus 94 Unternehmensmerkmalen verglichen und fanden, dass Entscheidungsbäume und neuronale Netze am besten abschneiden, mit großen ökonomischen Gewinnen für Anleger, die in manchen Fällen die Performance führender regressionsbasierter Strategien verdoppeln; die Gewinne stammen vor allem aus nichtlinearen Wechselwirkungen zwischen Merkmalen. Freyberger, Neuhierl und Weber (2020) kamen mit nichtparametrischen Verfahren zu einem ähnlichen Schluss, Bianchi, Büchner und Tamoni (2021) erweiterten die Evidenz auf Anleiherisikoprämien, und Kelly und Xiu (2023) legten eine umfassende Darstellung dessen vor, was sie Financial Machine Learning nennen. Zwei Einschränkungen sind für das Folgende wichtig. Israel, Kelly und

[5] Random Forests mitteln viele Entscheidungsbäume, die auf Bootstrap-Stichproben der Daten wachsen; Gradient Boosting passt Bäume nacheinander an die Residuen der vorherigen an. Beide verringern die Varianz eines einzelnen Baums ohne vorgegebene Funktionsform.

Moskowitz (2020) betonten, dass Finanzdaten im Vergleich zu jenen Daten, mit denen maschinelles Lernen anderswo Erfolge feiert, klein und verrauscht sind, sodass ökonomische Restriktionen unverzichtbar bleiben; Avramov, Cheng und Metzker (2023) zeigten, dass ein Großteil des scheinbaren Gewinns schwindet, sobald Kleinstwerte, notleidende Aktien oder Phasen hoher Marktvolatilität ausgeschlossen werden.[6]

Genetische Algorithmen, die in der Themenstellung dieser Arbeit genannt werden, gehören zur selben Familie. Von Holland (1975) eingeführt, durchsuchen sie einen großen Raum möglicher Regeln oder Portfoliogewichte durch Selektion, Rekombination und Mutation einer Population von Lösungen. Im Asset Management wurden sie zur Auswahl von Handelsregeln, zur Kalibrierung von Faktorgewichten und zur Lösung von Portfolioproblemen mit Nebenbedingungen eingesetzt, an denen konvexe Optimierer scheitern, etwa Kardinalitätsgrenzen oder Mindeststückzahlen. Ihre Flexibilität ist zugleich ihre Gefahr: Eine evolutionäre Suche über Tausende Regelvarianten ist ein Lehrbuchfall des Problems multipler Tests, das Kapitel 8.2 analysiert.

2.3 Generative KI und autonome Agenten

Generative Modelle unterscheiden sich von Prognosemodellen in drei ökonomisch bedeutsamen Hinsichten. Erstens sind ihre Eingaben unstrukturiert. Die Mitschrift eines Earnings Calls, der Risikobericht eines Geschäftsberichts, die Erklärung einer Notenbank oder ein Nachrichtenstrom sind keine Zahlenvektoren, sondern Satzfolgen, deren Bedeutung vom Kontext abhängt; Cao et al. (2023) zeigten, dass Unternehmen ihre Offenlegungen bereits für die Maschinen schreiben, die sie lesen. Zweitens ist ihre Ausgabe Inhalt und nicht eine Punktschätzung. Das Modell sagt nicht nur, dass der Tonfall eines Calls negativ ist; es kann erklären, warum, die konkrete Prognoseänderung herausarbeiten, die Notiz entwerfen, die ein Analyst geschrieben hätte, und sie zu einem Kundenbrief umarbeiten. Drittens wird ihr Verhalten angewiesen statt trainiert: Dasselbe Modell klassifiziert, extrahiert, fasst zusammen oder schließt je nach Eingabeaufforderung, was die Modellentwicklung teilweise zu einem Problem der Spezifikation und Prüfung von Anweisungen macht.[7]

Autonome Agenten erweitern diese Fähigkeiten um das Handeln. Ein Agent ist ein Sprachmodell, das ein Ziel, Zugang zu Datenquellen und Programmen sowie die Fähigkeit erhält, das Ziel in Schritte zu zerlegen, diese Programme aufzurufen, die Ergebnisse zu prüfen und über das weitere Vorgehen zu entscheiden. Im Anlagekontext kann ein Agent beauftragt werden, eine Beobachtungsliste zu überwachen, die neuesten Berichte

[6] Nach Avramov, Cheng und Metzker (2023) stammen die Gewinne von Machine-Learning-Signalen aus schwer arbitragierbaren Aktien und aus Marktphasen mit hohen Arbitragehemmnissen.

[7] Die Fähigkeit, neue Aufgaben aus Anweisungen zu erledigen, das sogenannte In-Context Learning, wurde von Brown et al. (2020) dokumentiert; sie macht ein einziges allgemeines Modell ohne Neutraining in vielen Arbeitsabläufen der Vermögensverwaltung einsetzbar.

abzurufen, gemeldete Zahlen mit dem Konsens zu vergleichen, ein Bewertungsmodell zu aktualisieren und eine Empfehlung mit Quellen zu entwerfen, alles ohne weitere menschliche Aufforderung. Balyasny Asset Management etwa berichtet von einer internen Research-Plattform mit spezialisierten Agenten, die Notenbankreden analysieren oder Merger-Arbitrage-Situationen beurteilen (OpenAI, 2026). Der Schritt von der Assistenz zur Autonomie ist zugleich jener Schritt, bei dem Fragen der Verantwortung, des Modellrisikos und der Regulierung akut werden; die Kapitel 7 und 8 kommen darauf zurück.[8]

Abbildung 1 stellt die beiden Familien entlang der typischen Abfolge von Schritten gegenüber. Im klassischen Prozess ist das Modell eine Stufe in einer Abfolge von Umformungen von strukturierten Daten zu einem optimierten Portfolio, und der Mensch prüft das Ergebnis am Ende. Im generativen Prozess liest das Modell unstrukturierte Quellen, erzeugt Signale und Inhalte, und ein Agent koordiniert weitere Schritte; der Mensch ist in die Schleife eingebunden, bevor die Ausgabe zur Entscheidung wird. Tabelle 1 fasst die Unterschiede entlang von acht Dimensionen zusammen.

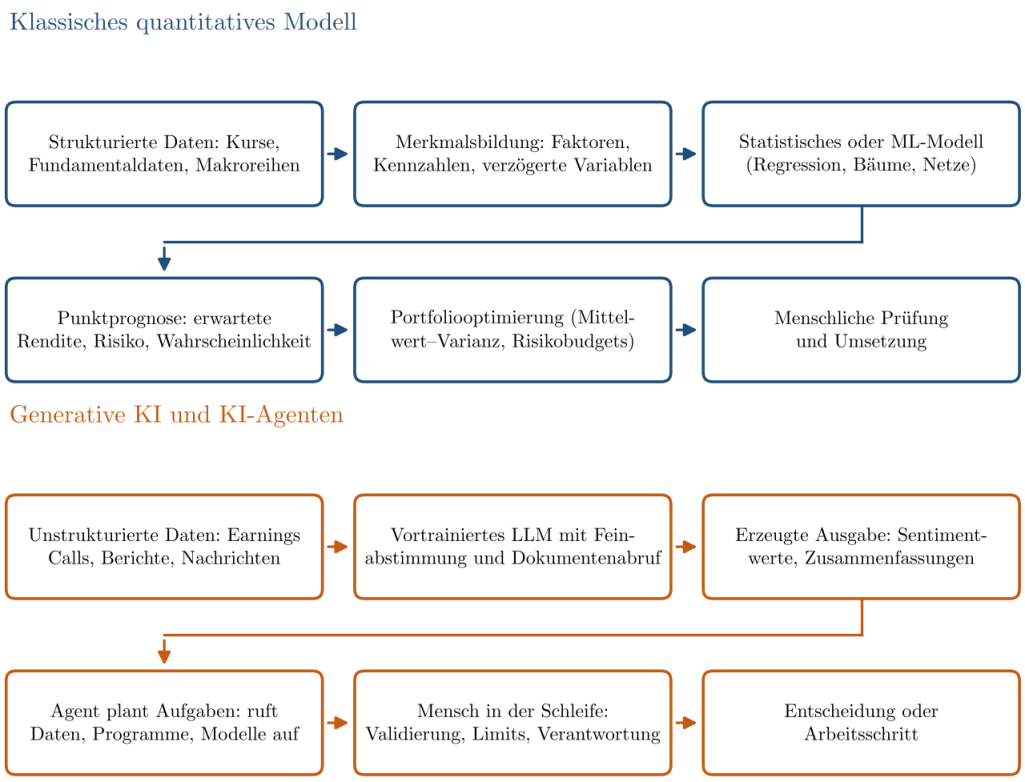


Abbildung 1: Klassischer quantitativer Prozess und generativer KI-Prozess im Asset Management
Quelle: Eigene Darstellung.

[8] Die Definition eines KI-Systems in Art. 3 Abs. 1 der KI-Verordnung verweist ausdrücklich auf „unterschiedliche Grade von Autonomie", sodass agentische Systeme unabhängig vom zugrunde liegenden Modell in den Anwendungsbereich der Verordnung fallen.

Tabelle 1: Klassische KI und generative KI im Asset Management im Vergleich

Dimension	Klassische KI / Prognosemodelle	Generative KI / LLMs und Agenten
Hauptzweck	Mustererkennung und Prognose (Renditen, Risiko, Ausfall)	Erzeugung von Inhalten (Zusammenfassungen, Analysen, Berichte) und Erledigung von Aufgaben
Typische Modelle	Regression, penalisierte Regression, Bäume, Boosting, neuronale Netze, genetische Algorithmen	Transformerbasierte Sprachmodelle (GPT, BERT, FinBERT, BloombergGPT), Systeme mit Dokumentenabruf, Agenten
Eingaben	Strukturierte Daten: Kurse, Fundamentaldaten, Makroreihen, alternative Daten	Unstrukturierte Daten: Mitschriften, Berichte, Nachrichten, Research, E-Mails; strukturierte Daten über Programme
Ausgaben	Punktprognosen, Wahrscheinlichkeiten, Scores	Text, strukturierte Extraktionen, Sentimentwerte, Entscheidungsentwürfe
Entwicklung	Training und Validierung mit historischen Daten	Vortraining durch einen Anbieter; Anweisung, Feinabstimmung und Dokumentenabruf durch den Nutzer
Bewertung	Out-of-Sample-Fehler, Sharpe-Ratio, Kalibrierung	Aufgabengenauigkeit, faktische Konsistenz, menschliche Prüfung; Look-ahead-Bias schwer auszuschließen
Typischer Fehler	Überanpassung, Scheinkorrelation, Regimewechsel	Halluzination, Inkonsistenz, Empfindlichkeit gegenüber der Anweisung, Datenleck aus dem Trainingskorpus
Ökonomische Wirkung	Erhöht die Präzision privater Signale	Senkt die Kosten der Verarbeitung öffentlicher Information; beschleunigt Arbeitsabläufe

Quelle: Eigene Darstellung in Anlehnung an Gu, Kelly und Xiu (2020), Kelly und Xiu (2023), Brown et al. (2020) und Wu et al. (2023).

2.4 Meilensteine der KI im Finanzwesen

Die Geschichte der KI im Finanzwesen ist länger, als die jüngste Debatte vermuten lässt, und Abbildung 2 reiht die Meilensteine. Die Dartmouth-Konferenz von 1956 gab dem Gebiet seinen Namen (McCarthy et al., 1955); die Wiederentdeckung der Backpropagation durch Rumelhart, Hinton und Williams (1986) machte mehrschichtige Netze trainierbar; White (1988) wandte sie auf die Tagesrenditen von IBM an und fand mit bewundernswerter Offenheit keine Evidenz gegen die Hypothese effizienter Märkte. Hochreiter und Schmidhuber (1997) führten das Long-Short-Term-Memory-Netz ein, das später die Modellierung von Sequenzen dominierte. Die quantitative Revolution im Asset Management verlief parallel, von Faktormodellen bis zur statistischen Arbitrage, und ihre Zerbrechlichkeit wurde im August 2007 sichtbar, als Khandani und Lo (2011) aus den

gleichzeitigen Verlusten vieler quantitativer Fonds schlossen, dass deren Strategien überfüllt waren.[9]

Die textliche Wende begann mit Tetlock (2007) und Loughran und McDonald (2011), die Deep-Learning-Wende mit dem Durchbruch in der Bilderkennung durch Krizhevsky, Sutskever und Hinton (2012) und dem Transformer 2017. Die ersten Fonds, die die Aktienauswahl vollständig an maschinelles Lernen delegierten, wurden 2017 aufgelegt, darunter der Acatis AI Global Equities in Deutschland und der AI Powered Equity ETF in den Vereinigten Staaten (private banking magazin, 2017).[10] Vortrainierte Sprachmodelle folgten 2018 und 2020, ChatGPT 2022, BloombergGPT und die ersten LLM-Studien zur Renditeprognose 2023. Die regulatorische Antwort kam mit der EU-KI-Verordnung, in Kraft seit 1. August 2024, und der öffentlichen Erklärung der Europäischen Wertpapier- und Marktaufsichtsbehörde zur KI in Wertpapierdienstleistungen (ESMA, 2024); der Digital Omnibus zur KI vom Juli 2026 verschob anschließend die Pflichten für Hochrisiko-Systeme (Europäische Union, 2026). Im selben Zeitraum legte Bridgewater einen Fonds von rund zwei Milliarden US-Dollar auf, der weitgehend von maschinellem Lernen geführt wird (Bloomberg, 2024), und agentische Anwendungen verbreiteten sich in der Branche.

[9] Khandani und Lo (2011) rekonstruieren die Ereignisse vom 6. bis 9. August 2007 mit einer einfachen Contrarian-Strategie und zeigen, dass auf Verluste beispielloser Größe eine ebenso rasche Umkehr folgte, was auf die erzwungene Auflösung ähnlicher Portfolios durch mehrere große Fonds hindeutet.

[10] Der AI Powered Equity ETF (Kürzel AIEQ) wurde im Oktober 2017 von EquBot aufgelegt und wählt US-Aktien mithilfe von IBM Watson aus (ThinkAdvisor, 2017); nach Rupprechter (2026) blieb er hinter dem S&P 500 zurück.

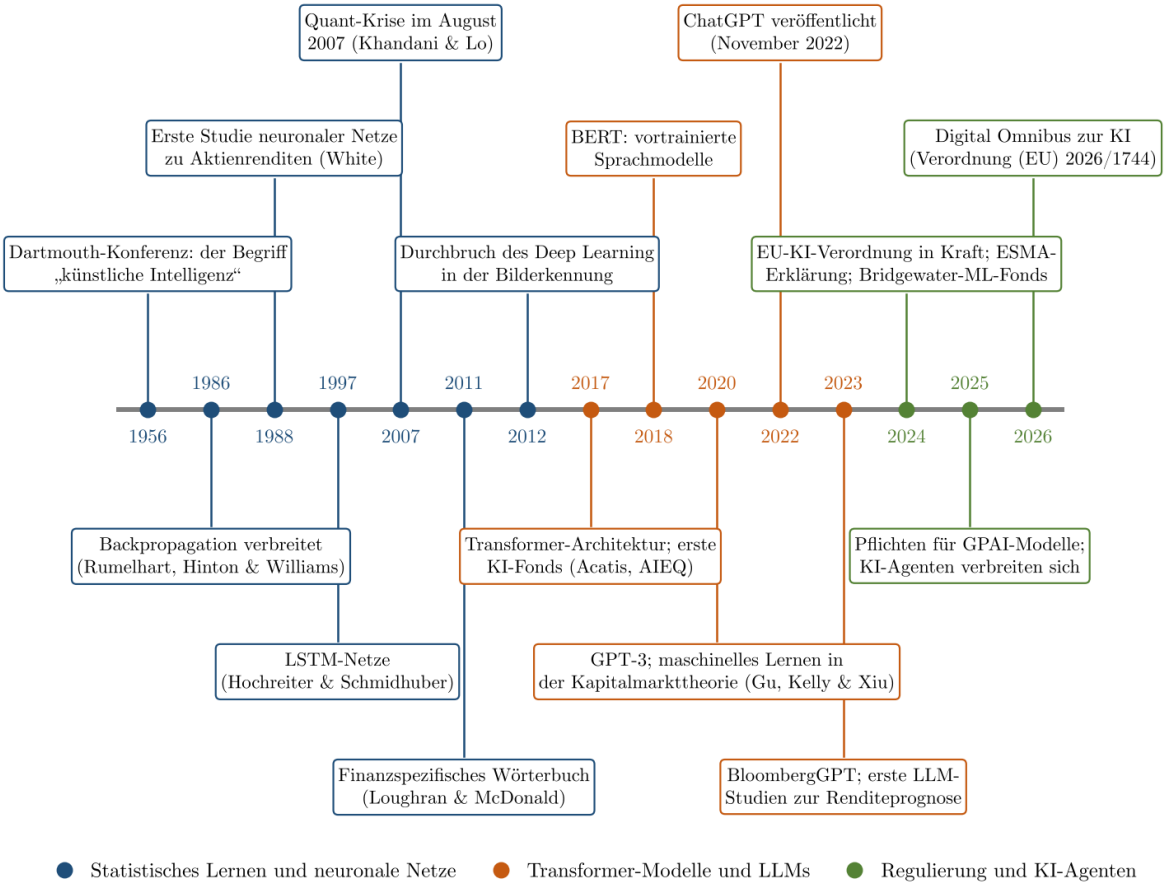


Abbildung 2: Meilensteine der künstlichen Intelligenz im Finanzwesen, 1956 bis 2026
Quelle: Eigene Darstellung auf Basis der im Text zitierten Quellen.

3 Informationseffizienz der Aktienmärkte im Zeitalter der KI

3.1 Die Effizienzmarkthypothese und das Grossman-Stiglitz-Paradoxon

Fama (1970) definierte einen Markt als effizient, wenn die Kurse die verfügbare Information vollständig widerspiegeln, und unterschied die schwache, die halbstrenge und die strenge Form danach, ob die Informationsmenge vergangene Kurse, alle öffentlichen Informationen oder alle Informationen einschließlich privater Signale umfasst. Die Hypothese war nie als Behauptung gemeint, die Kurse seien immer richtig; Fama (1991) selbst betonte, dass Effizienz nur gemeinsam mit einem Modell erwarteter Renditen geprüft werden kann.[11] Grossman und Stiglitz (1980) lieferten den Grund, warum Märkte nicht vollkommen effizient sein können: Würden die Kurse alle Information offenbaren, würde niemand dafür

[11] Fama (1991) ersetzte die drei Formen durch drei Familien von Tests, nämlich Tests auf Renditeprognostizierbarkeit, Ereignisstudien und Tests auf private Information; die Ereignisstudie in Kapitel 5 gehört zur zweiten Familie.

bezahlen, sie zu beschaffen, und beschaffte sie niemand, könnten die Kurse sie nicht offenbaren. Im Gleichgewicht verdienen informierte Händler auf ihre kostspielige Information gerade eine Entschädigung, und der Grad der Effizienz hängt von den Informationskosten und von jenem Rauschen ab, das die Kurse daran hindert, die Information vollständig preiszugeben. Kyle (1985) formalisierte, wie informierte Händler ihren Vorteil schrittweise ausnutzen, sodass Information über die Zeit und nicht auf einmal in die Kurse eingeht.

Künstliche Intelligenz tritt über die Kostenparameter in diesen Rahmen ein. Eine Technologie, die die Kosten der Beschaffung privater Information senkt, erhöht den Anteil informierter Händler und macht die Kurse informativer; eine Technologie, die die Kosten der Verarbeitung öffentlicher Information senkt, verkürzt die Verzögerung, mit der öffentliche Nachrichten eingepreist werden. Die Hypothese adaptiver Märkte von Lo (2004) fügt eine dynamische Dimension hinzu: Gewinnchancen werden durch Innovation geschaffen, von frühen Anwendern genutzt und mit der Verbreitung der Innovation wegkonkurriert, sodass Effizienz ein bewegliches Ziel und kein Zustand ist. Stein (2009) zeigte, dass anspruchsvolle Anleger die Effizienz sogar verringern können, wenn sie sich in denselben Strategien drängen und ihr Fremdkapitaleinsatz sie zu gleichzeitiger Auflösung zwingt, was der Mechanismus hinter den Ereignissen von 2007 ist.[12]

3.2 Verarbeitungskosten, begrenzte Aufmerksamkeit und textliche Information

Zwischen dem Eintreffen öffentlicher Information und ihrer Einpreisung liegt die Arbeit des Lesens. Hirshleifer und Teoh (2003) modellierten Anleger mit begrenzter Aufmerksamkeit, die auf hervorstechende Kennzahlen und nicht auf den detaillierten Inhalt von Offenlegungen reagieren, und Da, Engelberg und Gao (2011) maßen Aufmerksamkeit unmittelbar über Suchvolumina im Internet. Blankespoor, deHaan und Marinovic (2020) gliederten die Literatur nach dem Begriff der Verarbeitungskosten von Offenlegungen: den Kosten, von einer Offenlegung zu erfahren, sie zu beschaffen und ihren Inhalt in eine Bewertung zu integrieren. Ihre Übersicht argumentiert, dass diese Kosten ökonomisch relevant sind, dass sie zwischen Anlegern variieren und dass sie eine Erklärung für die verzögerte Reaktion auf Gewinnmeldungen bieten, die Ball und Brown (1968) sowie Bernard und Thomas (1989) als Erste dokumentierten.[13]

Die eindrucksvollste Evidenz für Verarbeitungskosten stammt aus Texten, die öffentlich, billig zu beschaffen und dennoch unbeachtet sind. Cohen, Malloy und Nguyen

[12] Stein (2009) nennt dies das Problem überfüllter Positionen und das Problem des Fremdkapitaleinsatzes; beide entstehen, weil ein einzelner Arbitrageur nicht beobachten kann, wie viele andere dieselbe Position halten, eine Beschränkung, die identische KI-Modelle verschärfen.

[13] Verarbeitungskosten unterscheiden sich von Beschaffungskosten: Ein Geschäftsbericht ist kostenlos abrufbar, doch seine Integration in eine Bewertung erfordert Stunden fachkundigen Lesens, und genau diese Kosten senken Sprachmodelle.

(2020) zeigten, dass Unternehmen, die die Sprache ihrer Berichte ändern, anschließend unterdurchschnittlich abschneiden: Es gibt keinen Ankündigungseffekt, und die Renditen fallen erst an, wenn die Information später durch Nachrichten, Ereignisse oder Gewinnmeldungen offenbar wird; ein Portfolio, das „Veränderer“ leerverkauft und „Nichtveränderer“ kauft, erzielte im Zeitraum 1995–2014 Alphas von bis zu 188 Basispunkten pro Monat. Fedyk und Hodson (2023) dokumentierten, dass Neukombinationen alter Information größere Kursbewegungen und anschließende Umkehrungen auslösen als unmittelbare Wiederholungen. Kim, Muhn und Nikolaev (2023) ließen Unternehmensberichte von ChatGPT zusammenfassen und fanden, dass deren „Aufblähung“ mit höherer Informationsasymmetrie zwischen Anlegern einhergeht. Diese Ergebnisse stecken das Gebiet ab, in dem generative KI die Effizienz verändern kann: Wo die bindende Beschränkung die Kosten des Lesens sind, wird eine Maschine, die billig liest, die Kurse näher an ihren Informationsgehalt heranführen.[14]

3.3 Datenfülle und Informationsgehalt der Kurse

Ob mehr Daten die Kurse informativer machen, ist theoretisch eine offene Frage. Farboodi und Veldkamp (2020) modellierten das langfristige Wachstum der Finanzdatentechnologie und zeigten, dass die Auswertung des Orderflusses anderer Anleger die Produktion fundamentaler Information verdrängt; in ihrem Modell beeinträchtigt dies den Informationsgehalt der Kurse nicht, verbessert aber auch nicht zwingend die Marktliquidität. Dugast und Foucault (2018) zeigten, dass Kurse billige, aber verrauschte Signale früher widerspiegeln und dass dies den Informationsgehalt der Kurse letztlich senken kann, weil es die Nachfrage nach präziseren Signalen verringert. Dessaint, Foucault und Frésard (2024) lieferten das empirische Gegenstück: Mit alternativen Daten verlagern Prognostiker ihre Aufmerksamkeit von der langen auf die kurze Frist, sodass ihre kurzfristigen Prognosen informativer und ihre langfristigen weniger informativ werden. Bai, Philippon und Savov (2016) fanden, dass der Informationsgehalt von US-Aktienkursen über die Gewinne in drei bis fünf Jahren seit 1960 zugenommen hat, und zwar vor allem bei Unternehmen mit höherem institutionellem Anteilsbesitz und Aktienumschlag, bei Unternehmen mit gehandelten Optionen und bei Wachstumsunternehmen.[15]

Diese Ergebnisse sind für generative KI unmittelbar relevant. Sprachmodelle sind außerordentlich gut für die nahe Zukunft: Sie arbeiten heraus, was ein Unternehmen gestern gesagt hat und worin es sich von dem unterscheidet, was es im Vorquartal sagte. Für die Bewertung von Zahlungsströmen in einem Jahrzehnt, die von Urteilen über

[14] Cohen, Malloy und Nguyen (2020) messen Änderungen mit Ähnlichkeitsmaßen zwischen aufeinanderfolgenden Berichten, etwa der Kosinus-Ähnlichkeit von Wortvektoren; ein Rückgang der Ähnlichkeit zeigt an, dass das Unternehmen seine Sprache geändert hat.

[15] Bai, Philippon und Savov (2016) messen den Informationsgehalt der Kurse über die Fähigkeit aktueller Marktbewertungen, künftige Gewinne vorherzusagen, nach Kontrolle der laufenden Gewinne.

Wettbewerb, Technologie und Politik abhängt, die kein Korpus enthält, sind sie weit weniger offensichtlich nützlich. Die natürliche Hypothese lautet daher, dass generative KI die kurzfristige Informationseffizienz erhöht, die Zeit verkürzt, in der Nachrichten in die Kurse eingehen, und die Drift verkürzt, während ihre Wirkung auf den langfristigen Informationsgehalt unbestimmt ist und negativ sein kann, wenn sie den Researchaufwand auf das lenkt, was gelesen werden kann, statt auf das, was verstanden werden muss.[16]

3.4 Hypothesen zur Wirkung generativer KI

Aus der Verbindung der Argumente formulieren wir vier Hypothesen, die das Forschungsdesign in Kapitel 5 und künftige Arbeiten prüfen können. Hypothese 1 (Geschwindigkeit): Der Anteil der gesamten Kursreaktion auf einen Earnings Call, der innerhalb des ersten Handelstags eintritt, ist seit der Verbreitung LLM-gestützter Analyse gestiegen, und die Drift nach der Ankündigung hat sich verkürzt. Hypothese 2 (Inhalt): Aus LLMs abgeleitete Sentiment-Maße erklären einen größeren Teil der Ankündigungsrendite als wörterbuchbasierte Maße, und die zusätzliche Erklärungskraft konzentriert sich auf qualitative Passagen wie die Frage-und-Antwort-Runde statt auf die vorbereiteten Ausführungen. Hypothese 3 (Crowding): Wenn dieselben Grundmodelle von vielen Anlegern genutzt werden, sinkt die Streuung ihrer Sentiment-Signale, die Korrelation der Transaktionen rund um Ankündigungen steigt, und die Rentabilität einer eigenständigen Sentiment-Strategie nimmt im Zeitverlauf ab, im Einklang mit Lo (2004) und Stein (2009). Hypothese 4 (Reflexivität): Emittenten passen ihre Sprache an die Modelle an, wie Cao et al. (2023) für die Maschinenlesbarkeit von 10-K-Berichten dokumentierten, sodass das gemessene Sentiment vorbereiteter Ausführungen weniger informativ wird, während das Sentiment unvorbereiteter Antworten informativ bleibt.

Zwei weitere Überlegungen prägen das erwartete Ausmaß. Die erste ist die Heterogenität der Anleger. Blue-Chip-Aktien werden von Dutzenden Analysten und von anspruchsvollen systematischen Anlegern verfolgt, sodass die Grenzwirkung einer zusätzlichen Lesemaschine kleiner ist als bei wenig beachteten Nebenwerten; die stärksten Effekte generativer KI auf die Effizienz sollten daher außerhalb des Blue-Chip-Universums auftreten, was selbst eine prüfbare Implikation ist.[17] Die zweite ist das Verhalten des Emittenten. Druz et al. (2020) zeigten, dass Änderungen des Tonfalls des Managements, insbesondere eine zunehmende Negativität („bleak tone changes“), niedrigere künftige Gewinne vorhersagen und dass Analysten und Anleger darauf zu schwach reagieren; ein Modell, das Tonfallüberraschungen gegenüber einer unternehmensspezifischen Basis misst,

[16] Dessaint, Foucault und Frésard (2024) beschreiben dies als Verlagerung der Aufmerksamkeit von Prognostikern auf die kurze Frist; in unserer Terminologie ist es eine Verkürzung des Anlagehorizonts, über den Research Mehrwert stiftet. Die Hypothese selbst ist eine eigene.

[17] Die Erwartung, dass sich die Effekte auf kleine, wenig beachtete Aktien konzentrieren, entspricht der Evidenz von Avramov, Cheng und Metzker (2023), wonach sich die Gewinne des maschinellen Lernens dort konzentrieren, wo die Arbitragehemmnisse hoch sind.

wird einen Effekt nach unserer Einschätzung daher eher entdecken als ein Modell, das den Tonfall absolut misst. Das Forschungsdesign in Kapitel 5 übernimmt diese Spezifikation.

4 Empirische Evidenz: Text, Sentiment und Aktienrenditen

4.1 Von Wörterbüchern zu Transformern

Die Messung des Sentiments in Finanztexten hat drei Generationen durchlaufen. Die erste zählte Wörter. Tetlock (2007) wandte das psychosoziale Wörterbuch der Harvard-Universität auf die Kolumne „Abreast of the Market" des Wall Street Journal an und fand, dass hoher Pessimismus Kursdruck mit anschließender Umkehr vorhersagt, ein Muster, das eher für Stimmung als für Information spricht. Loughran und McDonald (2011) zeigten, dass fast drei Viertel der von diesem Wörterbuch als negativ eingestuften Wörter im Finanzkontext nicht negativ sind, „liability" und „tax" sind die offensichtlichen Beispiele, und erstellten Wortlisten für negativen, positiven, unsicheren und prozessbezogenen Tonfall, die zum Standard des Feldes wurden. Jegadeesh und Wu (2013) ließen den Markt die Wörter selbst gewichten, indem sie Renditen auf Worthäufigkeiten regressierten, und García (2013) zeigte mit einem Jahrhundert von Kolumnen der New York Times, dass sich die Wirkung des Sentiments auf Renditen in Rezessionen konzentriert.[18]

Die zweite Generation lernte die Gewichte. Ke, Kelly und Xiu (2019) schlugen ein überwachtes Themenmodell vor, das stimmungsgeladene Wörter aus ihrem Zusammenhang mit Renditen erkennt, und zeigten, dass es Wörterbücher bei Dow-Jones-Nachrichten übertrifft. Frankel, Jennings und Lee (2022) verglichen Machine-Learning-Maße mit wörterbuchbasierten Maßen zu den Veröffentlichungsterminen von 10-K-Berichten und Telefonkonferenzen und fanden eine signifikant höhere Erklärungskraft für Renditen, wobei Regressionsbäume nach dem Random-Forest-Verfahren am besten abschnitten. Heston und Sinha (2017) verwendeten einen mit neuronalen Netzen ermittelten Sentimentwert und dokumentierten, dass tägliches Nachrichtensentiment Renditen für ein bis zwei Tage vorhersagt, während wöchentliche Aggregate Renditen für ein Quartal vorhersagen. Bybee et al. (2024) gewannen Themen aus dem Wall Street Journal und zeigten, dass die Aufmerksamkeit, die Wirtschaftsnachrichten einzelnen Themen widmen, die wirtschaftliche Aktivität eng nachzeichnet.[19]

Die dritte Generation liest. Vortrainierte Transformer stellen jedes Wort im Kontext seines Satzes und jeden Satz im Kontext seines Dokuments dar, wodurch sie Verneinung,

[18] García (2013) verwendet den Anteil positiver und negativer Wörter in zwei Finanzkolumnen der New York Times von 1905 bis 2005; in Rezessionen sagt ein Schock seines Nachrichtenmaßes um eine Standardabweichung eine Änderung der bedingten durchschnittlichen Rendite des Dow Jones Industrial Average um 12 Basispunkte über einen Tag voraus.

[19] Überwachte Sentimentmodelle benötigen Renditen als Lernsignal und laufen daher Gefahr, genau jenes Rauschen zu lernen, das sie messen sollen; ihr Wert lässt sich nur außerhalb der Stichprobe belegen.

Abschwächung und Ironie erkennen, die Wortzählungen nicht sehen können. Huang, Wang und Yang (2023) stimmten BERT auf Finanzkommunikation ab und berichteten, dass das so entstandene FinBERT das Sentiment von Sätzen aus Analystenberichten genauer klassifiziert als das Loughran-McDonald-Wörterbuch und als Naive Bayes, Support-Vektor-Maschinen, Random Forests sowie konvolutionale und LSTM-Netze, mit dem größten Vorteil bei kleinen Trainingsstichproben und bei Texten mit Finanzbegriffen, die in der Alltagssprache selten sind; andere Verfahren unterschätzen den textlichen Informationsgehalt von Earnings Calls gegenüber FinBERT um mindestens 18 Prozent. Wu et al. (2023) wiesen nach, dass ein mit Finanztexten vortrainiertes Modell allgemeine Modelle ähnlicher Größe bei Finanzaufgaben übertrifft und bei allgemeinen Aufgaben wettbewerbsfähig bleibt. Der Übergang vom Zählen zum Lesen ist nicht bloß technisch: Er verändert den Gegenstand der Messung vom Vokabular zur Bedeutung und macht die Messung von einem Modell abhängig, dessen Trainingsdaten der Forscher nicht kontrolliert.[20]

4.2 Earnings Calls als Informationsereignis

Telefonkonferenzen zu den Quartalsergebnissen sind das wichtigste wiederkehrende Offenlegungsereignis börsennotierter Unternehmen: Das Management präsentiert das Quartal, gibt einen Ausblick und beantwortet Fragen von Analysten. Der Call verbindet vorbereitete und unvorbereitete Kommunikation und ist daher ein idealer Gegenstand der Textanalyse. Price et al. (2012) maßen den Tonfall von Mitschriften der Calls mit einem finanzspezifischen Wörterbuch, das sie dem allgemeinen Harvard-IV-4-Wörterbuch überlegen fanden, und zeigten, dass der Tonfall des Calls ein signifikanter Prädiktor abnormaler Renditen und des Handelsvolumens ist und Gewinnüberraschungen über die 60 Handelstage nach dem Call dominiert; die Frage-und-Antwort-Runde besitzt zusätzliche Erklärungskraft für die Drift nach der Gewinnmeldung, konzentriert auf Unternehmen, die keine Dividende zahlen. Mayew und Venkatachalam (2012) gingen über Wörter hinaus und maßen den emotionalen Zustand der Manager aus der Tonaufnahme; stimmliche Hinweise sind informativ für die künftige Unternehmensleistung. Larcker und Zakolyukina (2012) stuften Calls auf Basis späterer Bilanzkorrekturen als täuschend ein und zeigten, dass linguistische Modelle der Ausführungen von CEO und CFO eine Zufallsklassifikation um 6 bis 16 Prozent übertreffen und mindestens so gut abschneiden wie Modelle auf Basis von Bilanzvariablen.[21]

[20] FinBERT ist als Modell mit offenen Gewichten verfügbar, was die Replikation von Studien ohne Zugang zu kommerziellen Schnittstellen und mit dokumentiertem Trainingskorpus erlaubt.

[21] In den Vereinigten Staaten verpflichtete die Regulation Fair Disclosure des Jahres 2000 (Securities and Exchange Commission, 2000) die Emittenten, wesentliche Informationen allen Anlegern gleichzeitig zugänglich zu machen, was den Earnings Call zu jener öffentlichen Aufzeichnung machte, auf der diese Literatur beruht.

Druz et al. (2020) verfeinerten die Messung, indem sie zeigten, dass Änderungen des Tonfalls des Managements Information tragen: Eine zunehmende Negativität („bleak tone changes“) sagt niedrigere künftige Gewinne und höhere Unsicherheit stark voraus, Analysten und Anleger reagieren darauf zu schwach, und eine darauf aufbauende Handelsstrategie erzielt abnormale Renditen. Cao et al. (2023) dokumentierten die andere Seite der Wechselwirkung: Mit der wachsenden maschinellen Leserschaft von Berichten mieden Unternehmen Wörter, die Algorithmen als negativ einstufen, und Manager von Unternehmen mit hoher maschineller Leserschaft zeigten sogar eine als positiver und erregter eingeschätzte Sprechemotion. Für die empirische Arbeit folgt daraus, dass Sentiment-Maße als Überraschungen gegenüber einer unternehmensspezifischen Basis zu spezifizieren sind und dass der Informationsgehalt vorbereiteter Passagen im Zeitverlauf abnehmen kann, wenn Emittenten lernen, für den Algorithmus zu schreiben.[22]

4.3 Evidenz aus großen Sprachmodellen

Die ersten Studien mit allgemeinen LLMs erschienen innerhalb von Monaten nach der Freigabe von ChatGPT. Lopez-Lira und Tang (2023) fragten das Modell, ob eine Schlagzeile gut oder schlecht für den Aktienkurs eines Unternehmens sei, und bildeten aus den Antworten Portfolios; mit Schlagzeilen, die nach dem Wissensstichtag der Modelle veröffentlicht wurden, fanden sie eine signifikante Prognostizierbarkeit der Renditen, wobei die Prognosefähigkeit im Allgemeinen mit der Modellgröße zunimmt. Chen, Kelly und Xiu (2022) bildeten mit LLMs wie ChatGPT und LLaMA Nachrichtenartikel aus 16 Aktienmärkten in 13 Sprachen ab und zeigten, dass diese Repräsentationen informativ für erwartete Renditen sind. Eine dritte Studie, die berichtete, GPT-4 sage die Richtung von Gewinnänderungen aus standardisierten Jahresabschlüssen besser voraus als Finanzanalysten (Kim, Muhn und Nikolaev, 2024), wurde im Februar 2025 von ihren Autoren zurückgezogen, nachdem ein Mitautor Unstimmigkeiten in Daten und Analysen festgestellt hatte; ihre Ergebnisse werden hier daher nicht verwendet, und die Episode selbst zeigt, wie fragil frühe Evidenz in einem sich rasch entwickelnden Feld sein kann.[23]

Jenseits der Renditen wurden LLMs eingesetzt, um zu lesen, was quantitative Modelle nicht lesen konnten. Jha et al. (2024) leiteten aus Mitschriften von Earnings Calls einen Score der vom Management erwarteten Änderungen der Investitionsausgaben ab, der die Investitionen bis zu neun Quartale im Voraus vorhersagt, und Unternehmen mit hohem Score erzielen anschließend negative abnormale Renditen. Hansen und Kazinnik (2023) zeigten, dass GPT-Modelle die geldpolitische Ausrichtung von FOMC-Mitteilungen

[22] Cao et al. (2023) messen die maschinelle Leserschaft näherungsweise über die erwarteten maschinellen Abrufe von Berichten aus dem EDGAR-System der SEC; die maschinellen Abrufe von 10-K- und 10-Q-Berichten stiegen von rund 361.000 im Jahr 2003 auf rund 165 Millionen im Jahr 2016.

[23] Die Studie von Lopez-Lira und Tang (2023) wurde seit April 2023 mehrfach überarbeitet; ihr Design, den Test auf Nachrichten nach dem Wissensstichtag der Modelle zu beschränken, ist das in Kapitel 5 übernommene Prinzip.

deutlich besser klassifizieren als andere gebräuchliche Verfahren und dass die Begründungen von GPT-4 menschlichem Urteil ebenbürtig sind. Eisfeldt, Schubert und Zhang (2023) konstruierten aus dem Aufgabeninhalt von Berufen ein Maß für die Exposition einzelner Unternehmen gegenüber generativer KI und fanden, dass stark exponierte Unternehmen in den zwei Wochen nach der Freigabe von ChatGPT höhere Renditen erzielten, ein erster Hinweis darauf, dass die Märkte die Technologie selbst bepreisen. Tabelle 2 fasst jene Studien zusammen, die für das Forschungsdesign in Kapitel 5 am wichtigsten sind.

Ein methodischer Vorbehalt betrifft den Look-ahead-Bias: Ein Modell, dessen Trainingskorpus über den Stichprobenzeitraum hinausreicht, kann sich „erinnern“, wie eine Geschichte ausging, statt es zu erschließen. Glasserman und Lin (2023) prüften dies, indem sie Unternehmensnamen aus Schlagzeilen entfernten, und fanden überraschend, dass die anonymisierten Schlagzeilen innerhalb des Trainingszeitraums besser abschnitten: Das allgemeine Wissen eines Modells über ein Unternehmen lenkt es stärker ab, als ein erinnertes Ergebnis ihm hilft, ein „Ablenkungseffekt“, der bei großen Unternehmen besonders ausgeprägt ist, während der Look-ahead-Bias außerhalb der Stichprobe kein Problem war. He et al. (2025) trainierten chronologisch konsistente Modelle mit Texten, die zum jeweiligen Zeitpunkt verfügbar waren, und kommen ebenfalls zu dem Schluss, dass der Look-ahead-Bias in einer Anwendung der Kapitalmarktforschung gering ist. Belastbare Designs schließen ihn dennoch konstruktiv aus: Sie beschränken den Testzeitraum auf Daten nach dem Trainingsende des Modells, anonymisieren Entitäten oder verwenden zeitpunktgenaue Modelle mit dokumentiertem Trainingskorpus. Jeff Shen von BlackRock Systematic hat den „Peek-ahead-Bias“ handelsüblicher Modelle als einen Hauptgrund dafür beschrieben, finanzspezifische Modelle im eigenen Haus zu bauen (Misonzhnik, 2025).

Tabelle 2: Ausgewählte empirische Studien zu textlicher Information, Sentiment und Aktienrenditen

Studie	Daten und Textquelle	Methode	Hauptergebnis
Tetlock (2007)	Kolumne des Wall Street Journal, 1984–1999	Harvard-Wörterbuch, VAR	Hoher Pessimismus sagt Kursdruck und Umkehr voraus
Loughran und McDonald (2011)	10-K-Berichte, 1994–2008	Finanzspezifische Wortlisten	Allgemeine Wörterbücher klassifizieren Finanzsprache falsch; neue Listen hängen mit
Price et al. (2012)	Mitschriften von Earnings Calls (USA)	Finanzspezifisches Wörterbuch vs. Harvard IV-4	Tonfall sagt abnormale Renditen, Volumen und 60-Tage-Drift voraus; Q&A ergänzt Information
Mayew und Venkatachalam (2012)	Tonaufnahmen von Earnings Calls	Analyse stimmlicher Emotion	Stimmliche Hinweise auf den Affekt des Managements sind informativ für die künftige
Jegadeesh und Wu (2013)	10-K-Berichte	Marktbasierte Wortgewichte	Renditegewichtete Wortwerte übertreffen gleichgewichtete Wörterbücher
Ke, Kelly und Xiu (2019)	Dow-Jones-Nachrichten	Überwachtes Sentiment-Themenmodell	Gelerntes Sentiment sagt Renditen besser voraus als Wörterbücher
Druz et al. (2020)	Mitschriften von Earnings Calls	Änderungen des Tonfalls des Managements	Zunehmende Negativität sagt niedrigere Gewinne voraus; Analysten und Anleger reagieren
Cohen, Malloy und Nguyen (2020)	10-K- und 10-Q-Berichte, 1995–2014	Dokumentähnlichkeit	Sprachänderungen in Berichten sagen Underperformance voraus; kein Ankündigungseffekt
Frankel, Jennings und Lee (2022)	10-K-Berichte, Telefonkonferenzen	Maschinelles Lernen (Random Forest am besten)	Gelerntes Sentiment erklärt Renditen besser als Wörterbuchmaße
Huang, Wang und Yang (2023)	Analystenberichte, Earnings Calls	FinBERT (abgestimmter Transformer)	Höhere Klassifikationsgenauigkeit als das LM-Wörterbuch und andere ML-Verfahren
Cao et al. (2023)	10-K-Berichte, EDGAR-Abrufe	Maß für maschinelle Leserschaft	Unternehmen meiden Wörter, die Algorithmen als negativ lesen; positiverer Stimmklang
Lopez-Lira und Tang (2023)	Schlagzeilen nach dem Wissensstichtag der Modelle	GPT-Anweisungen, Portfolios	Signifikante Prognostizierbarkeit; nimmt im Allgemeinen mit der Modellgröße zu
Chen, Kelly und Xiu (2022)	Nachrichten in 16 Märkten, 13 Sprachen	LLM-Repräsentationen (z. B. ChatGPT, LLaMA)	LLM-basierte Repräsentationen sind informativ für erwartete Renditen
Glasserman und Lin (2023)	Originale vs. anonymisierte Schlagzeilen	GPT-Sentiment mit maskierten Entitäten	In der Stichprobe schneiden anonymisierte Schlagzeilen besser ab (Ablenkungseffekt); außerhalb
Jha et al. (2024)	Mitschriften von Earnings Calls	ChatGPT-Investitionsscore	Score sagt Investitionsausgaben bis zu neun Quartale voraus
He et al. (2025)	Zeitpunktgenaue Textkorpora	Chronologisch konsistente LLMs	Look-ahead-Bias in einem Kapitalmarkttest gering

Quelle: Eigene Darstellung auf Basis der zitierten Studien.

5 Forschungsdesign: KI-gestützte Sentiment-Analyse und Kursreaktionen auf Earnings Calls von Blue-Chip-Aktien

5.1 Forschungsfrage und Hypothesen

Die Beispielfrage, die diese Arbeit veranlasst hat, lautet: Welchen Einfluss hat KI-gestützte Sentiment-Analyse auf Kursreaktionen bei Earnings Calls, empirisch untersucht für Blue-Chip-Aktien? Die Frage hat zwei Ebenen. Die erste ist, ob mit einem LLM gemessenes Sentiment Information über die Kursreaktion enthält, die über die quantitative Gewinnüberraschung und über den wörterbuchbasierten Tonfall hinausgeht, also Hypothese 2 aus Kapitel 3.4. Die zweite ist, ob die Verbreitung solcher Hilfsmittel unter Anlegern die Dynamik der Reaktion selbst verändert hat, also Hypothese 1, die Geschwindigkeitshypothese, und Hypothese 3, die Crowding-Hypothese. Das hier beschriebene Design behandelt beide Ebenen mit einem einzigen Datensatz und folgt der Tradition der Ereignisstudie von Ball und Brown (1968), Brown und Warner (1985) und MacKinlay (1997), wie sie Kothari und Warner (2007) zusammenfassen.

5.2 Stichprobe und Daten

Blue-Chip-Aktien sind aus drei Gründen das angemessene Universum für einen ersten Test. Ihre Mitschriften sind vollständig und über lange Zeiträume verfügbar; ihre Kurse sind liquide, sodass abnormale Renditen nicht durch Mikrostrukturrauschen verzerrt werden; und sie sind genau jene Aktien, bei denen der kleinste Effekt zu erwarten ist, sodass ein signifikantes Ergebnis aussagekräftig wäre. Das Universum umfasst die Mitglieder des EURO STOXX 50, des DAX und des S&P 100[24] zum jeweiligen Quartalsende über die 46 Quartale vom ersten Quartal 2015 bis zum zweiten Quartal 2026. Da der DAX mit Wirkung vom 20. September 2021 von 30 auf 40 Werte erweitert wurde (Deutsche Börse, 2021) und deutsche Unternehmen, die sowohl dem DAX als auch dem EURO STOXX 50 angehören, nur einmal gezählt werden, umfasst das Universum bis zum zweiten Quartal 2021 rund 165 und danach rund 175 Unternehmen je Quartalsende. Nach eigener Schätzung ergibt dies höchstens rund 7.800 Unternehmensquartale und damit Earnings Calls ($26 \times 165 + 20 \times 175 = 7.790$); europäische Unternehmen, die Calls nur halbjährlich abhalten, verringern die Zahl weiter.[25] Weil sich die Indexmitgliedschaften ändern, liegt die Zahl

[24] Der S&P 100 statt des S&P 500 wird verwendet, um die transatlantische Stichprobe ausgewogen zu halten und sich auf die am stärksten beachteten Unternehmen zu konzentrieren, für die der Test am konservativsten ist.

[25] Eigene Berechnung in vier Schritten: (i) Indexplätze $50 + 30 + 100 = 180$ bis zum zweiten Quartal 2021 und $50 + 40 + 100 = 190$ ab dem dritten Quartal 2021; (ii) abzüglich angenommener 15 deutscher Unternehmen, die zum jeweiligen Quartalsende sowohl dem DAX als auch dem EURO STOXX 50 angehören, ergibt rund 165 bzw. rund 175 Unternehmen, wobei sich der S&P 100 mit den europäischen Indizes nicht überschneidet; (iii) 26 Quartale vom ersten Quartal 2015 bis zum zweiten Quartal 2021 und 20 Quartale vom dritten Quartal 2021 bis zum zweiten Quartal 2026; (iv) $26 \times 165 + 20 \times 175 = 7.790$ Unternehmensquartale als Obergrenze für die Zahl der Calls. Die

verschiedener Unternehmen über den Gesamtzeitraum über 175; maßgeblich für die Ereignisstudie ist die Zahl der Ereignisse. Entscheidend ist, dass die Stichprobe einen Zeitraum vor und einen Zeitraum nach der Verbreitung generativer KI umfasst.

Drei Datenquellen sind erforderlich. Mitschriften von Earnings Calls, getrennt in vorbereitete Ausführungen und Frage-und-Antwort-Runde und mit minutengenauem Zeitstempel, sind bei kommerziellen Anbietern wie LSEG Refinitiv oder S&P Capital IQ verfügbar. Tageskurse, Renditen und Marktkapitalisierungen stammen von denselben Anbietern oder, für US-Aktien, von CRSP; Faktorrenditen für Europa und die Vereinigten Staaten stellt die Datenbibliothek von Kenneth French bereit. Analystenkonsensschätzungen vor dem Call und das gemeldete Ergebnis je Aktie, aus denen die standardisierte unerwartete Gewinnkomponente berechnet wird, stammen von I/B/E/S. Der Zeitstempel ist wichtig: Viele europäische Calls finden während der Handelszeit statt, und der Ereignistag ist als die erste Handelssitzung zu definieren, in der auf den vollständigen Call gehandelt werden konnte, was Intraday-Daten oder eine konservative Konvention erfordert.[26]

Die Stichprobe wird in üblicher Weise bereinigt. Calls, die mit anderen wesentlichen Ankündigungen zusammenfallen, etwa Fusionen, Gewinnwarnungen oder Dividendenkürzungen, werden gekennzeichnet und in einem Robustheitstest ausgeschlossen; Unternehmen mit weniger als acht vorangegangenen Calls werden ausgeschlossen, damit eine unternehmensspezifische Tonfallbasis geschätzt werden kann; und Schätzfenster, die sich mit dem vorherigen Call überschneiden, werden verkürzt. Tabelle 3 fasst die Designelemente und ihre Begründung zusammen.

5.3 Messung des Sentiments

Das Sentiment wird auf drei Arten gemessen, damit der zusätzliche Beitrag generativer KI isoliert werden kann. Das Wörterbuchmaß folgt Loughran und McDonald (2011): Der Nettotonfall ist die Differenz zwischen positiven und negativen Wörtern, geteilt durch die Gesamtzahl der Wörter, getrennt berechnet für die vorbereiteten Ausführungen und die Frage-und-Antwort-Runde. Das Maß des abgestimmten Transformers folgt Huang, Wang und Yang (2023): Jeder Satz wird von FinBERT als positiv, neutral oder negativ klassifiziert, und der Dokumentwert ist der Anteil positiver abzüglich des Anteils negativer Sätze. Das LLM-Maß folgt der strukturierten Anweisung von Lopez-Lira und Tang (2023), vom Verfasser dieser Arbeit erweitert: Das Modell erhält die anonymisierte Mitschrift, aus der Unternehmensnamen, Kürzel und Daten entfernt wurden, und wird angewiesen, auf einer Skala von minus fünf bis plus fünf die Auswirkungen auf die Gewinne des

Zahl der Doppelmitglieder ist eine Annahme; sie schwankt mit Indexänderungen von Quartalsende zu Quartalsende und ist bei Durchführung der Studie aus den historischen Indexzusammensetzungen zu bestimmen.

[26] Die Datenbibliothek von Kenneth R. French stellt tägliche Faktorrenditen für die Vereinigten Staaten und für Europa bereit, sodass abnormale Renditen über beide Regionen hinweg konsistent berechnet werden können.

Unternehmens in den nächsten vier Quartalen zu bewerten, die Zuversicht des Managements einzuschätzen und die drei Aussagen zu nennen, die die Bewertung am stärksten beeinflusst haben. Jede Mitschrift wird bei einer Stichprobentemperatur von null mehrfach bewertet, und der Median wird verwendet.

Drei Vorkehrungen begegnen den in Kapitel 4.3 erörterten Verzerrungen. Das Modell muss ein dokumentiertes Trainingsende vor Beginn des Testzeitraums haben, oder es ist ein Modell mit offenen Gewichten und veröffentlichtem Korpus zu verwenden, sodass der Testzeitraum außerhalb der Trainingsdaten liegt. Die Maskierung von Entitäten wie bei Glasserman und Lin (2023) beseitigt sowohl das Risiko, dass sich das Modell erinnert, was einem Unternehmen widerfuhr, als auch die Ablenkung durch sein Vorwissen über das Unternehmen, die nach ihrer Evidenz bei großen Unternehmen besonders ausgeprägt ist, also gerade bei jenen einer Blue-Chip-Stichprobe. Schließlich werden alle drei Maße nach Druz et al. (2020) als Überraschungen ausgedrückt: der Wert des aktuellen Calls abzüglich des Durchschnittswerts der vorangegangenen acht Calls desselben Unternehmens, geteilt durch die Standardabweichung dieser Werte. Die Tonfallüberraschung ist die interessierende Variable.[27]

5.4 Methodik der Ereignisstudie

Abnormale Renditen werden relativ zu einem Faktormodell berechnet, das über ein Fenster von 120 Handelstagen geschätzt wird, das 20 Tage vor dem Call endet. Die Basis ist das Marktmodell von Brown und Warner (1985); das Dreifaktorenmodell von Fama und French (1993) und das Vierfaktorenmodell von Carhart (1997) dienen als Robustheitsprüfungen. Für das Unternehmen i und den Ereignistag t ergibt sich die abnormale Rendite aus Gleichung (1)[28]

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

mit den über das Schätzfenster geschätzten Parametern α_i und β_i und der Rendite des Marktindex $\mathrm{R}_{m,t}$, und die kumulierte abnormale Rendite über das Ereignisfenster von τ_1 bis τ_2 aus Gleichung (2)

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

Gleichung (2) wird für drei Fenster ausgewertet: das Ankündigungsfenster von Tag 0 bis Tag +1, das Driftfenster von Tag +2 bis Tag +30 und das Gesamtfenster von Tag 0 bis Tag +30. Die statistische Signifikanz durchschnittlicher abnormaler Renditen wird mit dem standardisierten Querschnittstest von Boehmer, Musumeci und Poulsen (1991)

[27] Die Standardisierung beseitigt auch Skalenunterschiede zwischen den drei Maßen, sodass die Koeffizienten über die Maße hinweg in Einheiten einer unternehmensspezifischen Standardabweichung vergleichbar sind.

[28] Der Abstand von 20 Handelstagen zwischen dem Ende des Schätzfensters und dem Ereignis stellt sicher, dass die Vorwegnahme der Ankündigung die geschätzten Parameter nicht verzerrt.

beurteilt, der gegenüber ereignisbedingter Varianz robust ist, sowie mit der Anpassung von Kolari und Pynnönen (2010) für Querschnittskorrelation, die von Bedeutung ist, weil sich Calls in der Berichtssaison zeitlich häufen. Abbildung 3 veranschaulicht das erwartete Muster in stilisierter Form: ein Sprung am Ereignistag in Richtung der Sentiment-Überraschung, gefolgt von einer Drift in dieselbe Richtung, die das Kennzeichen einer unvollständigen sofortigen Reaktion ist.

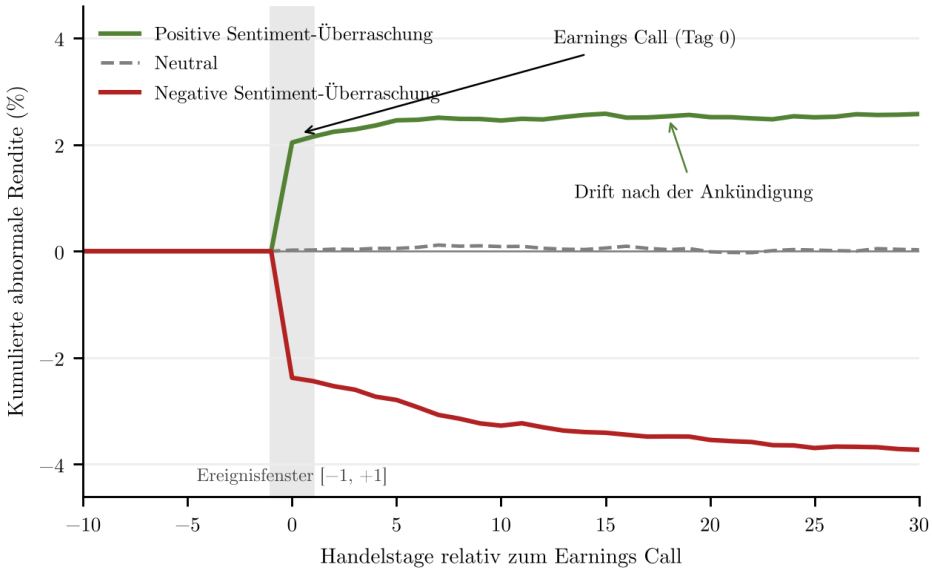


Abbildung 3: Stilisierte kumulierte abnormale Renditen rund um einen Earnings Call bei positiver, neutraler und negativer Sentiment-Überraschung
Anmerkungen: Schematische Darstellung des hypothetischen Musters; die Verläufe sind nicht aus Daten geschätzt und dienen nur der Erläuterung des Designs.
Quelle: Eigene Darstellung.

Die zweite Stufe setzt die kumulierten abnormalen Renditen aus Gleichung (2) in der Querschnittsregression der Gleichung (3) mit den Sentiment-Überraschungen in Beziehung,

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

wobei SUE die standardisierte unerwartete Gewinnkomponente ist, X ein Vektor von Kontrollvariablen aus Größe, Buch-Markt-Verhältnis, Momentum des Vorjahres, Analystenabdeckung, einem Indikator für Prognoseänderungen und dem Stand des VIX-Index am Ereignistag, und die Standardfehler nach Petersen (2009) nach Unternehmen und Quartal geclustert werden. Hypothese 2 sagt $\gamma_1 > 0$ in Gleichung (3) voraus, wobei die LLM-Überraschung ihre Signifikanz behält, wenn die Wörterbuchüberraschung einbezogen wird. Hypothese 1 wird geprüft, indem die Überraschungen mit einem Indikator für Calls nach dem ersten Quartal 2023 interagiert werden und indem das Verhältnis der Reaktion im Ankündigungsfenster zur Reaktion im Gesamtfenster vor und nach diesem Zeitpunkt verglichen wird; ein steigendes Verhältnis und ein schrumpfender

Driftkoeffizient, also γ_1 in Gleichung (3), geschätzt für das Driftfenster, zeigen eine schnellere Einpreisung an. Hypothese 3 wird geprüft, indem die Querschnittsstreuung der LLM-Werte über Modellversionen hinweg und das abnormale Handelsvolumen rund um Calls mit dem Verbreitungsindikator in Beziehung gesetzt werden.[29]

Tabelle 3: Design der vorgeschlagenen Ereignisstudie zu KI-gestütztem Sentiment und Reaktionen auf Earnings Calls

Element	Spezifikation	Begründung und Referenz
Universum	Mitglieder von EURO STOXX 50, DAX (bis Q2 2021 30, danach 40 Werte) und S&P 100 zum Quartalsende; rund 165 bzw. 175 Unternehmen	Liquide Blue Chips; vollständige Mitschriften; konservativer Test
Zeitraum	Q1 2015 bis Q2 2026 (46 Quartale); höchstens rund 7.800 Calls	Umfasst die Zeiträume vor und nach der Verbreitung generativer KI
Ereignis	Earnings Call mit Zeitstempel; Tag 0 = erste vollständige Handelssitzung	Price et al. (2012); Intraday-Zeitpunkt europäischer Calls
Texteinheiten	Vorbereitete Ausführungen und Q&A getrennt	Q&A ergänzt Information zur Drift (Price et al., 2012)
Sentiment-Maße	Loughran-McDonald-Nettotonfall; FinBERT-Satzwerte; LLM-Bewertung der anonymisierten Mitschrift	Loughran und McDonald (2011); Huang, Wang und Yang (2023); Lopez-Lira und Tang (2023)
Überraschungsdefinition	Wert abzüglich Unternehmensmittel der vorangegangenen acht Calls, standardisiert	Tonfalländerungen sind informativ (Druz et al., 2020)
Schutz vor Look-ahead	Trainingsende vor Testzeitraum; Maskierung von Entitäten; zeitpunktgenaues Modell	Glasserman und Lin (2023)
Abnormale Renditen	Marktmodell; FF3 und Carhart als Robustheit; Schätzfenster 120 Tage, Ende $t-20$	Brown und Warner (1985); MacKinlay (1997)
Fenster	$[0, +1]$, $[+2, +30]$, $[0, +30]$	Sofortige Reaktion versus Drift (Bernard und Thomas, 1989)
Tests	Standardisierter BMP-Test; Kolari-Pynnönen-Anpassung; geclusterte Regressionen	Boehmer, Musumeci und Poulsen (1991); Kolari und Pynnönen (2010); Petersen (2009)
Kontrollvariablen	SUE, Größe, Buch-Markt-Verhältnis, Momentum, Abdeckung, Prognoseänderung, VIX	Isoliert qualitative Information von der quantitativen Überraschung
Strukturbruch	Interaktion mit Indikator nach Q1 2023; Driftverhältnis	Hypothese 1 (Geschwindigkeit); Lo (2004)

Quelle: Eigene Darstellung.

[29] Die Verbreitung LLM-gestützter Hilfsmittel unter Anlegern ist nicht unmittelbar beobachtbar; als Näherungen dienen die Zahl maschineller Abrufe von Mitschriften, Befragungsdaten wie Mercer (2024) und die seit der Freigabe von ChatGPT verstrichene Zeit.

5.5 Erwartete Ergebnisse, Fallstricke und Robustheit

Die bisherige Literatur erlaubt eine begründete, wenn auch qualitative Erwartung. Price et al. (2012) und Druz et al. (2020) dokumentieren ökonomisch bedeutsame Effekte des Tonfalls von Calls auf Renditen in breiten US-Stichproben; angesichts der von Huang, Wang und Yang (2023) berichteten höheren Klassifikationsgenauigkeit sollten FinBERT- und LLM-Maße die Wörterbuchmaße mindestens erreichen. Bei Blue Chips werden die Effekte nach unserer Einschätzung kleiner sein als in jenen Studien; dies ist eine Erwartung, kein Ergebnis. Der Driftkoeffizient, also γ_1 in Gleichung (3) für das Driftfenster, ist die interessantere Größe: Gilt Hypothese 1, so sollte die mit LLM-gemessenen Überraschungen verbundene Drift nach 2023 zurückgegangen sein, selbst wenn die Ankündigungsreaktion es nicht ist, was ein unmittelbarer Beleg dafür wäre, dass generative KI die Informationseffizienz von Blue-Chip-Kursen erhöht hat.

Die Fallstricke sind jene jeder textbasierten Forschung, verschärft durch die Undurchsichtigkeit der Grundmodelle. Multiple Tests über Anweisungen, Modelle, Fenster und Teilstichproben laden zu jener Überanpassung ein, die Kapitel 8.2 analysiert; das Design legt die Anweisung und die Fenster daher fest, bevor die Daten untersucht werden, und berichtet jede gerechnete Spezifikation. Der Empfindlichkeit gegenüber der Anweisung wird begegnet, indem die Bewertung mit umformulierten Anweisungen wiederholt und die Korrelation der resultierenden Werte berichtet wird. Der Modelldrift wird begegnet, indem die Modellversion festgeschrieben und die Rohausgaben archiviert werden. Der umgekehrten Kausalität, der Möglichkeit, dass der Tonfall des Managements auf den Kurstrend vor dem Call reagiert, wird begegnet, indem der Kursanstieg vor dem Ereignis kontrolliert wird und die Überraschungsspezifikation verwendet wird. Schließlich ist die ökonomische Bedeutung jedes Ergebnisses nach Transaktionskosten und mit Blick auf die Kapazität einer Strategie zu beurteilen, nach dem Protokoll von Arnott, Harvey und Markowitz (2019); ein statistisch signifikanter Tonfalleffekt von zehn Basispunkten bei Standardwerten ist ein wissenschaftliches Ergebnis über Effizienz, nicht notwendigerweise eine profitable Strategie.[30]

6 Anwendungen I: Fondsmanagement

6.1 Signalgewinnung und Alpha-Research

In Publikumsfonds ist der älteste und nach wie vor wichtigste Einsatz der KI die Gewinnung von Renditeprognosen, von Signalen, für die Titelauswahl. Systematische Manager verbinden Hunderte Signale aus Kursen, Fundamentaldaten, Analystenverhalten

[30] Nach der Datenpolitik der Working-Paper-Reihe des Autors würden bei einer empirischen Umsetzung alle Kursdaten und jeder Rechenschritt in einer begleitenden Excel-Datei dokumentiert, wie es in dieser Arbeit für die Berechnungen hinter Abbildung 5 geschieht.

und alternativen Daten; BlackRock Systematic, mit rund 336 Milliarden US-Dollar verwaltetem Vermögen und einem Team von 230 Personen, berichtet, dass die Zahl der Signale in seiner US-Aktienstrategie von drei bei der Auflegung 1985 auf mehr als tausend gestiegen ist (Misonzhnik, 2025). Maschinelles Lernen setzt an zwei Punkten an: bei der Bildung einzelner Signale aus Rohdaten, etwa eines Sentimentwerts aus Mitschriften oder einer Nachfrageschätzung aus Stellenanzeigen, und bei der Verbindung von Signalen zu einer Prognose, wo nichtlineare Modelle der von Gu, Kelly und Xiu (2020) bewerteten Art die lineare Gewichtung ersetzen. Kelly, Malamud und Zhou (2024) haben argumentiert, dass sehr hochdimensionale Modelle entgegen der Intuition die Out-of-Sample-Leistung verbessern können, selbst wenn sie die Trainingsdaten interpolieren, sofern sie angemessen regularisiert werden; das Argument stützt den Einsatz komplexer Modelle, nicht aber den Verzicht auf ökonomische Vorannahmen.[31]

Generative KI verändert den ersten dieser Punkte stärker als den zweiten. Ein Sprachmodell verwandelt einen Earnings Call in eine Menge strukturierter Variablen, die Änderung des Ausblicks, den Tonfall der Antworten, die Zahl ausweichender Antworten, die Themen, auf die Analysten gedrängt haben, die wie jedes andere Signal in dieselben Kombinationsmodelle eingespeist werden können. BlackRock Systematic beschreibt den Einsatz von LLMs zum Lesen alternativer Daten, sozialer Medien und regulatorischer Berichte in der Rolle eines Finanzanalysten und betont, dass allgemeine Modelle an das Finanzwesen angepasst werden müssen, um den Peek-ahead-Bias zu vermeiden (Misonzhnik, 2025). Der Anlagehorizont solcher Signale ist kurz, typischerweise einige Monate, was der in Kapitel 3.3 erörterten Stärke der Sprachmodelle für die nahe Zukunft entspricht.[32]

6.2 Portfoliokonstruktion und Risikomanagement

Die Portfoliokonstruktion ist seit Markowitz (1952) die Domäne der Optimierung, und die praktischen Schwierigkeiten der Mittelwert-Varianz-Optimierung, ihre Empfindlichkeit gegenüber Schätzfehlern in erwarteten Renditen und Kovarianzen, sind wohlbekannt. Black und Litterman (1992) begegneten dem ersten Problem, indem sie die erwarteten Renditen am Marktgleichgewicht verankerten und Anlegermeinungen einmischten; KI tritt auf natürliche Weise als Quelle solcher Meinungen auf, wobei die einer Meinung beigemessene Zuversicht aus dem Out-of-Sample-Leistungsnachweis des Modells abgeleitet wird. López de Prado (2016) begegnete dem zweiten Problem mit Hierarchical Risk Parity, einem Verfahren des maschinellen Lernens, das Anlagen nach Korrelation gruppiert und

[31] Ein Signal ist eine Variable, die im Modell des Managers mit künftigen relativen Renditen zusammenhängt; die Verbindung vieler schwach informativer Signale ist das Wesen des systematischen Investierens.

[32] Misonzhnik (2025) zitiert für die Signale von BlackRock Systematic einen Prognosezeitraum von drei bis vier Monaten und merkt an, dass die Gruppe bereits vor der Freigabe von ChatGPT Erkenntnisse aus der maschinellen Sprachverarbeitung nutzte.

das Risiko entlang des entstehenden Baums verteilt, ohne die Kovarianzmatrix zu invertieren; er berichtete eine bessere Out-of-Sample-Leistung als Minimum-Varianz-Portfolios. Cong et al. (2021) gingen weiter und trainierten ein tiefes Modell des verstärkenden Lernens, das Portfolios unmittelbar aus Unternehmensmerkmalen bildet und den Prognoseschritt ganz umgeht.[33]

Das Risikomanagement nutzt maschinelles Lernen zur Schätzung von Kovarianzstrukturen, zur Erkennung von Regimewechseln und zur Identifikation überfüllter Positionen. Generative KI ergänzt die Überwachung qualitativer Risiken: Ein Agent kann jeden neuen Bericht jedes Portfoliounternehmens lesen, den Risikobericht im Sinne von Cohen, Malloy und Nguyen (2020) mit der Vorversion vergleichen und den Manager auf Änderungen hinweisen, die ein menschlicher Leser aus Zeitgründen nicht gefunden hätte. Die quantitativen und die generativen Fähigkeiten ergänzen einander, statt zu konkurrieren: Erstere schätzen, wie viel das Portfolio verlieren könnte, Letztere erklären, was sich in der Welt verändert hat, das einen solchen Verlust auslösen könnte.

6.3 Research und Datenfilterung mit Natural Language Processing

Der zweite in der Themenstellung dieser Arbeit genannte Anwendungsfall, die schnelle Auswertung großer unstrukturierter Datenmengen, ist jener, in dem sich generative KI am schnellsten verbreitet hat. Ein Fundamentalanalyst, der dreißig Unternehmen abdeckt, erhält in einem typischen Quartal dreißig Mitschriften, mehrere hundert Brokerstudien, Tausende Nachrichtenmeldungen und einen Strom regulatorischer Berichte. Sprachmodelle fassen dieses Material zusammen, extrahieren und vergleichen es, beantworten Fragen dazu mit Verweisen auf die Quellstellen und entwerfen die erste Fassung der Analystennotiz. Wellington Management berichtet von internen Anwendungen, die das eigene Research-Archiv, Mitschriften und öffentliche Kommentare abfragen (Correlation One, 2026), und die Branchenbefragung von Mercer (2024) ergab, dass 54 Prozent der globalen Investmentmanager KI bereits in ihrer Anlagestrategie oder im Research zu Anlageklassen einsetzten und weitere 37 Prozent dies planten.[34]

Der ökonomische Wert dieses Einsatzes liegt weniger im Alpha als in Produktivität und Abdeckung. Brynjolfsson, Li und Raymond (2025) maßen bei Mitarbeitern im Kundenservice einen durchschnittlichen Produktivitätsgewinn von rund 14 bis 15 Prozent (je nach Fassung der Studie) durch einen generativen KI-Assistenten, mit den größten Gewinnen bei neuen und gering qualifizierten Beschäftigten, Noy und Zhang (2023) fanden ähnliche Gewinne beim professionellen Schreiben, und Dell'Acqua et al. (2023) zeigten bei Unternehmensberatern, dass die Gewinne bei Aufgaben innerhalb der Kompetenz des

[33] Hierarchical Risk Parity benötigt weder die Inversion noch die positive Definitheit der Kovarianzmatrix, weshalb das Verfahren stabil bleibt, wenn die Zahl der Anlagen die Zahl der Beobachtungen übersteigt.

[34] Die Analystennotiz bleibt nach der treuhänderischen Logik von Kapitel 7.4 ein menschliches Produkt; der erzeugte Entwurf verkürzt die Zeit bis zur ersten Fassung, nicht die Verantwortung für die endgültige.

Modells groß sind und außerhalb negativ werden können, was sie als gezackte Grenze bezeichnen. Für Research-Abteilungen folgt daraus, dass generative KI die Zahl der Unternehmen und Dokumente erweitert, die ein Analyst verfolgen kann, während die Qualität des Urteils in den schwierigen Fällen eine menschliche Variable bleibt.

6.4 Handel, Betrieb und Kundenkommunikation

Im Handel wird maschinelles Lernen seit Langem eingesetzt, um kurzfristige Liquidität vorherzusagen und Ausführungsstrategien zu wählen, und verstärkendes Lernen dient der Minimierung der Marktbeeinflussung. Der Flash Crash vom 6. Mai 2010, den Kirilenko et al. (2017) analysierten, bleibt die Mahnung, dass zusammenwirkende Algorithmen eine Dynamik erzeugen können, die keiner von ihnen beabsichtigt hat. Dou, Goldstein und Ji (2025) zeigten in Simulationen, dass Händler auf Basis des verstärkenden Lernens stillschweigend kolludieren lernen können, ohne zu kommunizieren, was die Kurseffizienz verringert; das Ergebnis entspricht den Preisexperimenten von Calvano et al. (2020) und ist eine der systemischen Sorgen von Kapitel 8.4. Im Betrieb entwirft generative KI Fondsberichte, Kommentare und regulatorische Dokumente, gleicht Daten ab und beantwortet interne Anfragen; diese Anwendungen sind arm an Glanz und reich an messbarem Nutzen.[35]

6.5 Praktische Beispiele

Tabelle 4 versammelt dokumentierte Beispiele. Sie reichen von der vollständig autonomen Aktienauswahl bis zur Assistenz. Am einen Ende delegiert der im Juli 2017 aufgelegte Acatis AI Global Equities die Aktienauswahl und die Portfoliozusammensetzung vollständig an Deep-Learning-Modelle, die mit Fundamentaldaten bis zurück ins Jahr 1986 trainiert wurden, wobei der Fondsmanager nur die Umsetzung sicherstellt (private banking magazin, 2017); nach fast acht Jahren verwaltete der Fonds rund 37 Millionen Euro, und die Fachpresse berichtete über drei Jahre eine Underperformance gegenüber dem MSCI World, die mit der Neigung des Modells zu kleinen und mittleren Werten zusammenhing (Thieringer, 2025). Bridgewater legte 2024 einen Fonds von rund zwei Milliarden US-Dollar auf, dessen Entscheidungen weitgehend von maschinellem Lernen einschließlich LLMs mehrerer Anbieter getroffen werden (Bloomberg, 2024). Am anderen Ende stattete Morgan Stanley (2024) seine Finanzberater mit Assistenten auf Basis von OpenAI-Modellen für den Zugang zum hauseigenen Research und für Gesprächszusammenfassungen aus, Goldman Sachs begann im Januar 2025, seinen GS AI Assistant für rund 10.000 Beschäftigte einzuführen, mit dem Ziel, binnen Jahresfrist alle Wissensarbeiter zu erreichen

[35] Dou, Goldstein und Ji (2025) zeigen, dass dem kollusiven Ergebnis zwei getrennte Mechanismen zugrunde liegen: Preisauslöserstrategien und eine Lernverzerrung, die sie als Over-Pruning bezeichnen; keiner von beiden erfordert eine Kommunikation zwischen den Algorithmen.

(CNBC, 2025), und JPMorgan Chase (2024) berichtete von mehr als 2.000 Fachleuten für KI und maschinelles Lernen und Datenwissenschaftlern sowie über 400 produktiven Anwendungsfällen. Im deutschsprachigen Markt kündigte Union Investment (2024) für Dezember 2024 eine der ersten fondsbasierten Vermögensverwaltungen auf KI-Basis an.[36]

Zwei Lehren ergeben sich aus den Beispielen. Erstens gibt es das vollständig autonome Modell seit fast einem Jahrzehnt, und es hat das Assistenzmodell nicht verdrängt; die von KI-geführten Fonds eingesammelten Vermögen bleiben im Verhältnis zur Branche klein, und ihre Leistung war unauffällig. Zweitens setzen die größten und anspruchsvollsten Manager KI durchgängig ein, halten aber Menschen für das Portfolio verantwortlich, und Abis (2020) fand schon vor der generativen Welle Unterschiede zwischen quantitativen und diskretionären Investmentfonds: Quantitative Fonds halten deutlich mehr Aktien, handeln etwas mehr und spezialisieren sich auf die Titelauswahl, während diskretionäre Fonds zwischen Titelauswahl und Markttiming wechseln und in Rezessionen besser abschneiden; die Performance quantitativer Fonds in Aufschwüngen hat im Zeitverlauf abgenommen, was sie auf Überfüllung zurückführt. Generative KI mag den Abstand in der Anpassungsfähigkeit verringern, doch sie beseitigt nicht die Notwendigkeit des Urteils darüber, was neuartig ist.

[36] Die eigene Praxis des Autors liegt im Bereich der Assistenz: Bei R&B Wealth Management unterstützt KI das Markt- und Titelresearch, während Allokations- und Selektionsentscheidungen von den Portfoliomanagern innerhalb der Anlagerichtlinien des Fonds getroffen werden. Zur Arbeitsteilung zwischen Mensch und Maschine siehe Rupprechter (2026).

Tabelle 4: Dokumentierte Anwendungen von KI und generativer KI im Fonds- und Asset Management

Haus	Anwendung	Jahr	Quelle
Acatis (Deutschland)	AI Global Equities: Aktienauswahl und Portfoliozusammensetzung vollständig durch Deep Learning auf Fundamentaldaten; rund 37 Millionen Euro nach acht Jahren	2017	private banking magazin (2017); Thieringer (2025)
BlackRock Systematic	Mehr als 1.000 Signale; LLMs lesen Berichte, Nachrichten und alternative Daten; NLP-Signale schon vor ChatGPT im Einsatz	seit ca. 2022	Misonzhnik (2025)
Bridgewater Associates	Fonds von rund 2 Milliarden US-Dollar, weitgehend von maschinellem Lernen einschließlich mehrerer LLMs geführt	2024	Bloomberg (2024)
Balyasny Asset Management	Interne KI-Research-Plattform mit spezialisierten Agenten für Notenbankreden und Merger Arbitrage	2022–2026	OpenAI (2026)
Wellington Management	Interne Anwendungen, die Research-Archiv, Mitschriften und öffentliche Kommentare abfragen	2025	Correlation One (2026)
JPMorgan Chase	Mehr als 2.000 KI/ML-Fachleute und Datenwissenschaftler, über 400 produktive Anwendungsfälle; Markenanmeldung IndexGPT für KI-gestützte Titelauswahl	2023–2024	CNBC (2023); JPMorgan Chase (2024)
Morgan Stanley Wealth Management	AI @ Morgan Stanley Assistant auf OpenAI-Basis, vollständig eingeführt im September 2023; Gesprächsnotizen-Tool Debrief, Juni 2024	2023–2024	Morgan Stanley (2024)
Goldman Sachs	GS AI Assistant, zunächst für rund 10.000 Beschäftigte; unternehmensweite Einführung geplant	2025	CNBC (2025)
Union Investment (Deutschland)	Fondsbasierte Vermögensverwaltung auf KI-Basis, angekündigt für Dezember 2024	2024	Union Investment (2024)

Anmerkungen: Angaben wie in den Quellen genannt; Fallstudien von Anbietern sind an ihrer Quelle erkennbar. *Quelle*: Eigene Darstellung.

7 Anwendungen II: Asset Management für Einzelmandate

7.1 Anlagerichtlinien, Eignung und Überwachung von Beschränkungen

Ein Einzelmandat unterscheidet sich von einem Fonds dadurch, dass das Portfolio auf einen Kunden zugeschnitten ist: auf seine Ziele, seinen Anlagehorizont, seine Risikotoleranz, seine steuerliche Situation, seinen Liquiditätsbedarf und seine persönlichen Beschränkungen. Nach Artikel 25 der MiFID II muss der Verwalter die Eignung der Dienstleistung für jeden Kunden beurteilen und das Portfolio jederzeit innerhalb der vereinbarten Richtlinien halten (Europäische Union, 2014). In der Praxis sind diese Richtlinien schriftliche, oft individuell verhandelte Dokumente, deren Regeln in Prüfungen

umzusetzen sind, die vor und nach jeder Transaktion laufen. Generative KI eignet sich für diese Aufgabe ungewöhnlich gut: Ein Sprachmodell kann den Mandatsvertrag, die Anlagerichtlinie und die vom Kunden genannten Ausschlüsse lesen, die Regeln in eine strukturierte Form bringen, und ein Agent kann sie auf das Portfolio und auf jede vorgeschlagene Order anwenden und Verstöße zur menschlichen Bestätigung kennzeichnen. Ausschlüsse, die vom Lesen abhängen, etwa eine Obergrenze für Umsätze aus einer umstrittenen Tätigkeit, lassen sich überwachen, indem das Modell die Nachhaltigkeitsberichte der Portfoliounternehmen liest, eine Aufgabe, die bisher teure Daten Dritter erforderte.[37]

Dieselbe Fähigkeit dient der Eignungsbeurteilung selbst. Die Erklärung der ESMA zum Einsatz von KI in Wertpapierdienstleistungen für Kleinanleger (ESMA, 2024) erinnert die Unternehmen daran, dass die Pflichten der MiFID II unabhängig von der eingesetzten Technologie gelten, dass das Leitungsorgan für KI-gestützte Entscheidungen verantwortlich bleibt und dass die Unternehmen im besten Interesse der Kunden handeln, ihre Systeme testen und überwachen, ihr Personal schulen und die Kunden informieren müssen, wenn KI in der Interaktion mit ihnen eingesetzt wird. Ein Modell, das den Eignungsbericht aus dem Kundenfragebogen, den Gesprächsnotizen und der vorgeschlagenen Allokation entwirft, verringert den Verwaltungsaufwand des Beraters; es verlagert nicht die Verantwortung für das Urteil.[38]

7.2 Personalisiertes Reporting und Kundenkommunikation

Der sichtbarste Nutzen generativer KI in Einzelmandaten ist die Kommunikation. Ein Quartalsbrief, der in der Sprache und auf dem Kenntnisstand des Kunden erklärt, was im Portfolio geschehen ist, warum der Verwalter so gehandelt hat und was dies für die Ziele des Kunden bedeutet, war früher ein knappes Gut, das den größten Mandaten vorbehalten blieb. Sprachmodelle erzeugen solche Briefe aus der Performance-Attribution, der Transaktionsliste und der Hausmeinung zu vernachlässigbaren Grenzkosten, und sie können dies für jeden Kunden und in jeder Sprache. Morgan Stanley (2024) berichtet, dass seine Berater ein Hilfsmittel auf Basis von OpenAI-Modellen namens Debrief nutzen, das mit Zustimmung des Kunden Gesprächsnotizen erstellt, Handlungspunkte hervorhebt und eine Nachfass-E-Mail entwirft, die der Berater bearbeitet; schon zuvor hatte das Haus einen

[37] Die Extraktion von Regeln aus Verträgen ist eine Aufgabe der Informationsgewinnung, bei der Transformer-Modelle hohe Genauigkeit erreichen; dennoch muss jede extrahierte Regel von der verantwortlichen Person bestätigt werden, bevor sie durchgesetzt wird.

[38] ESMA (2024) richtet sich an Unternehmen, die Wertpapierdienstleistungen für Kleinanleger erbringen, und schließt ausdrücklich die Nutzung von KI-Anwendungen Dritter durch Mitarbeiter ein; sie ist eine Erinnerung an bestehende Pflichten, keine neue Regel.

Assistenten eingeführt, der den Beratern raschen Zugang zum hauseigenen Research und Wissen verschafft.[39]

Das Risiko dieses Einsatzes ist faktischer Natur. Ein Brief, der eine falsche Rendite nennt, eine Position erfindet oder einen Verlust falsch zurechnet, beschädigt das Vertrauen unwiederbringlich, und die Neigung von Sprachmodellen, flüssige, aber falsche Aussagen zu erzeugen, die Ji et al. (2023) dokumentierten, ist keine seltene Fehlfunktion, sondern eine Eigenschaft der Technologie. Die Abhilfe liegt in der Architektur: Jede Zahl in einem erzeugten Brief muss aus der geprüften Reportingdatenbank stammen und eingefügt, nicht erzeugt werden, und jede qualitative Aussage muss auf eine Quellstelle zurückführbar sein. Häuser, die diese Disziplin einhalten, erhalten Berichte, die genauer sind als handgeschriebene; Häuser, die das Modell frei schreiben lassen, erhalten elegante Fiktion.

7.3 Steuerbewusste Optimierung und Rebalancing

Einzelmandate erlauben eine Optimierung auf der Ebene einzelner Kauftranchen, die Publikumsfonds nicht bieten können. Rebalancing mit Rücksicht auf realisierte Gewinne und Verluste, die Realisierung von Verlusten innerhalb der Grenzen des nationalen Steuerrechts und der Aufbau personalisierter Indexportfolios, bekannt als Direct Indexing, sind Rechenprobleme, die maschinelles Lernen gut bewältigt: Die Zielfunktion verbindet Tracking Error, erwartete Rendite, Transaktionskosten und Steuerlast, und die Nebenbedingungen sind für jeden Kunden spezifisch. Hier leistet die klassische Optimierung und nicht die generative KI die Arbeit, doch die generative Ebene macht das Ergebnis nutzbar: Sie erklärt dem Kunden und dem Berater, warum die vorgeschlagenen Transaktionen gewählt wurden und was die Alternativen gekostet hätten.[40]

7.4 KI-Assistenten für Berater und die treuhänderische Rolle des Menschen

Der Berater, der ein Einzelmandat führt, erfüllt zwei Funktionen, die Maschinen nicht erfüllen: die Ausübung des treuhänderischen Urteils und die Begleitung des Kundenverhaltens. Die erste ist eine rechtliche und ethische Stellung; der Verwalter ist dafür verantwortlich, im besten Interesse des Kunden zu handeln, und weder MiFID II noch die KI-Verordnung erlauben es, diese Verantwortung an ein System zu delegieren. Die zweite ist verhaltensbezogen. Kahneman (2011) und die Literatur zum

[39] Die periodische Berichterstattung an Kunden der Portfolioverwaltung ist nach Artikel 60 der Delegierten Verordnung (EU) 2017/565 (Europäische Union, 2017) mindestens vierteljährlich vorgeschrieben, was dem erzeugten Brief neben der geschäftlichen auch eine regulatorische Funktion gibt.

[40] In Österreich führt die depotführende Stelle den Verlustausgleich für sämtliche Depots durch, die der Steuerpflichtige bei ihr hält (§ 93 Abs. 6 EStG 1988; ausgenommen sind betriebliche, treuhändig gehaltene und Depots mit mehreren Inhabern); ein Ausgleich über mehrere Banken erfordert die Steuererklärung. Jede steuerbewusste Optimierung muss daher die tatsächliche Depotstruktur des Kunden abbilden.

Anlegerverhalten, die Barberis, Shleifer und Vishny (1998) sowie Daniel, Hirshleifer und Subrahmanyam (1998) formalisierten, beschreiben Anleger, die auf Nachrichten überreagieren, jüngste Renditen fortschreiben und Strategien im ungünstigsten Moment aufgeben; in der vorangegangenen Arbeit des Autors (Rupprechter, 2026) wurde die Fähigkeit, einen Kunden durch einen Kursrückgang hindurch an einer schriftlichen Strategie festzuhalten, als einer jener Bereiche herausgearbeitet, in denen der menschliche Berater einen klaren Vorteil behält. Ein KI-Assistent, der die Fakten aufbereitet, die Szenarioanalyse entwirft und das Gespräch dokumentiert, stärkt diese Funktion; ein Chatbot, der die bange Frage des Kunden um Mitternacht mit einer flüssigen, aber unbegründeten Beruhigung beantwortet, untergräbt sie.[41]

Die angemessene Arbeitsteilung folgt der Logik von Abbildung 7 in Kapitel 8.6: Je geringer die Folgen eines Fehlers und je leichter seine Entdeckung, desto mehr Autonomie darf das System haben. Die Zusammenfassung eines Gesprächs steht am einen Ende; die Entscheidung, die Aktienquote eines Pensionisten aufzulösen, am anderen. Tabelle 5 zeigt einen stilisierten Arbeitsablauf für ein Einzelmandat mit der Rolle der KI und der Rolle des Menschen in jedem Schritt.

[41] Daniel, Hirshleifer und Subrahmanyam (1998) führen Überreaktion auf Selbstüberschätzung und auf die Neigung zurück, Erfolge sich selbst zuzuschreiben; beides wird verstärkt, wenn eine flüssig formulierende Maschine die Vorannahme des Anlegers bestätigt.

Tabelle 5: Stilisierter Arbeitsablauf eines Einzelmandats mit den Rollen der KI und des Menschen

Schritt	Rolle der KI	Rolle des Menschen	Kontrolle
Aufnahme und Eignung	Gewinnt Kundendaten aus Dokumenten; entwirft Eignungsbeurteilung und Anlagerichtlinie	Führt das Gespräch; beurteilt Ziele und Risikotragfähigkeit; unterzeichnet die Beurteilung	Vier-Augen-Prüfung; Aufzeichnungspflichten der MiFID II
Strategische Allokation	Szenarioanalyse; Kapitalmarktannahmen; Entwurf des Vorschlags	Legt die Allokation innerhalb von Hausmeinung und Kundenvorgaben fest	Anlageausschuss
Titelauswahl	Liest Berichte und Calls; erzeugt Sentiment- und Qualitätswerte; prüft Ausschlüsse	Wählt Titel aus; weicht mit dokumentierten Gründen ab	Modellvalidierung; Protokoll der Abweichungen
Ordererzeugung	Schlägt Rebalancing-Transaktionen mit Steuer- und Kostenoptimierung vor	Gibt Orders frei	Richtlinienprüfung vor dem Handel
Überwachung	Laufende Richtlinien- und Limitüberwachung; liest Nachrichten zu Positionen; Warnhinweise	Entscheidet über Warnhinweise	Compliance nach dem Handel; Eskalationsregeln
Reporting	Entwirft personalisierte Briefe aus geprüften Daten; fasst Gespräche zusammen	Prüft und unterzeichnet; erklärt persönlich	Verankerung aller Zahlen; Prüfpfad
Überprüfung	Vergleicht Ergebnis mit Zielen; kennzeichnet Abweichungen von der Strategie	Hält den Kunden an der Strategie; passt bei veränderten Umständen an	Jährliche Eignungsüberprüfung

Quelle: Eigene Darstellung.

7.5 Fragen, die Kunden von Einzelmandaten stellen sollten

Für Kunden lautet die Frage nicht, ob ihr Verwalter KI einsetzt, was in wenigen Jahren alle Verwalter tun werden, sondern wie. Tabelle 6 nennt die Fragen, die ein Kunde stellen sollte, und die Warnzeichen in den Antworten. Die Fragen gelten gleichermaßen für den Anleger, der einen Fonds in Betracht zieht, der mit seinem KI-Einsatz wirbt: Die Existenz eines Modells ist kein Beleg für seinen Wert, und ein Backtest, der nicht vorab festgelegt wurde, der keine Transaktionskosten ausweist und der sich nicht ökonomisch erklären lässt, ist aus den Gründen, die Kapitel 8.2 entwickelt, als Ergebnis einer Suche und nicht als Entdeckung zu behandeln.

Tabelle 6: Fragen zur Beurteilung des KI-Einsatzes eines Vermögensverwalters und Warnzeichen

Kriterium	Frage an den Verwalter	Warnzeichen
Rolle der KI	Welche Entscheidungen trifft das System allein, und welche trifft eine Person?	Der Verwalter kann die Entscheidungen nicht nennen, für die ein Mensch verantwortlich bleibt
Ökonomische Begründung	Warum sollte das Signal funktionieren, und wer steht auf der anderen Seite des Geschäfts?	Die Erklärung verweist nur auf Genauigkeitskennzahlen, nicht auf einen ökonomischen Mechanismus
Belege	Ist der Leistungsnachweis real oder zurückgerechnet, nach Kosten, und über welchen Zeitraum?	Nur Backtests; kein Zeitraum außerhalb der Stichprobe; keine Transaktionskosten
Multiple Tests	Wie viele Modellvarianten wurden ausprobiert, bevor diese gewählt wurde?	Die Frage wird nicht verstanden, oder die Zahl ist nicht festgehalten
Daten	Könnte das Modell beim Training die Zukunft gesehen haben (Look-ahead-Bias)?	Allgemeines Modell auf historischen Zeiträumen ohne Maskierung von Entitäten oder Kontrolle des Trainingsendes
Validierung	Wer validiert das Modell unabhängig, und wie oft?	Die Entwickler validieren ihr eigenes Modell
Verankerung	Werden Zahlen in Berichten vom Modell erzeugt oder aus geprüften Daten übernommen?	Der Verwalter kann es nicht sagen
Abhängigkeit	Von welchen Drittanbietern hängt das System ab, und was geschieht bei deren Ausfall?	Ein einziger externer Modellanbieter ohne Ersatzlösung
Regulierung	Wie werden die Pflichten aus MiFID II, KI-Verordnung und DORA erfüllt?	Kein Programm zur KI-Kompetenz; kein Verzeichnis der KI-Systeme

Quelle: Eigene Darstellung in Anlehnung an Arnott, Harvey und Markowitz (2019), ESMA (2024) und IOSCO (2021).

8 Herausforderungen: messbarer Nutzen, Scheinkorrelationen, Modellrisiko und Regulierung

8.1 Messbarer Nutzen versus Hype

Jede Basistechnologie hat eine Phase durchlaufen, in der die Investitionen dem messbaren Ertrag vorausliefen. Solow (1987) bemerkte, das Computerzeitalter sei überall sichtbar außer in den Produktivitätsstatistiken, und Brynjolfsson, Rock und Syverson (2021) erklärten das Muster mit einer Produktivitäts-J-Kurve: Unternehmen müssen zunächst immaterielle Ergänzungen aufbauen, neue Prozesse, Daten, Fähigkeiten und Organisationsformen, bevor sich die Technologie auszahlt, sodass die gemessene Produktivität zunächst sinkt. Generative KI steht am Anfang dieser Kurve. McKinsey & Company (2023) schätzte das Potenzial generativer KI auf 2,6 bis 4,4 Billionen US-Dollar jährlicher Wertschöpfung über alle Branchen, mit dem Bankwesen unter den größten

Nutznießern; die in Abbildung 4 gesammelte Befragungsevidenz zeigt, wie weit die Praxis vom Potenzial noch entfernt ist.

Die Studie von KPMG (2026) zum gegenwärtigen Stand der KI im deutschen Asset Management, gestützt auf Experteninterviews mit Verantwortlichen aus dem COO-Ressort von Asset Managern, Servicing-Anbietern und Verwahrstellen sowie auf mehr als 640 Einzelbewertungen, ergab, dass alle befragten Häuser KI als strategisch wichtig einstufen und 75 Prozent sie als kritisch oder existenzentscheidend für ihre Wettbewerbsfähigkeit ansehen, dass aber 63 Prozent noch in Test- oder Proof-of-Concept-Phasen stecken und nur 12 Prozent KI erfolgreich skaliert haben und einen messbaren Return on Investment erzielen. In den Worten der Studie kann nur eine kleine Minderheit einen messbaren wirtschaftlichen Nutzen erzielen, und der Nutzen wächst mit der Tiefe der Integration in die Wertschöpfungskette, nicht mit der Zahl der Anwendungsfälle. Die Bank of England und die Financial Conduct Authority (2024) stellten fest, dass 75 Prozent der britischen Finanzdienstleister KI bereits einsetzten und weitere 10 Prozent dies innerhalb von drei Jahren planten, und Mercer (2024) fand, dass 54 Prozent der globalen Investmentmanager KI in ihrer Anlagestrategie oder im Research zu Anlageklassen nutzten; in einer Folgebefragung von 131 Managern Anfang 2026 meldeten 55 Prozent KI in mindestens einem Investmentprozess, 27 Prozent befanden sich im Pilotstadium (Mercer, 2026). Die Verbreitung ist mit anderen Worten nahezu universell; der messbare Nutzen ist es nicht.[42]

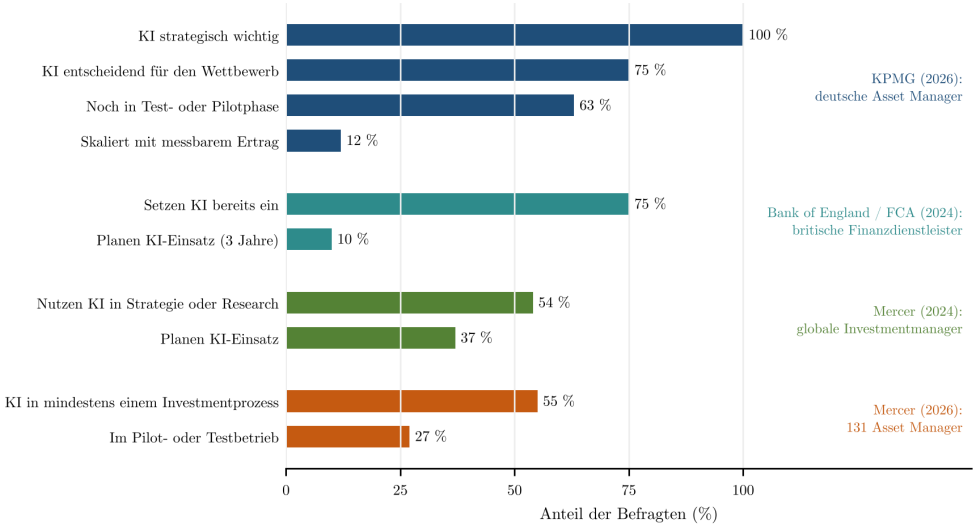


Abbildung 4: Verbreitung von KI und messbarer Nutzen im Asset Management: Befragungsevidenz
Quelle: Eigene Darstellung auf Basis von KPMG (2026), Bank of England und Financial Conduct Authority (2024), Mercer (2024) sowie Mercer (2026).

Drei Gründe erklären die Lücke gerade bei Investmentgesellschaften. Der erste ist das Problem der kleinen Daten nach Israel, Kelly und Moskowitz (2020): Das Verhältnis von

[42] KPMG (2026) führt die Lücke auf eine Führung zurück, die Anwendungsfälle und Nutzungszahlen zählt statt Geschäftsergebnisse; erfolgreiche Häuser verankern KI nach dieser Darstellung in der Wertschöpfungskette und führen sie über Geschäftsziele.

Signal zu Rauschen in Renditen ist gering und instabil, sodass die spektakuläre Genauigkeit von Sprachmodellen bei sprachlichen Aufgaben nicht auf die Prognose übertragbar ist. Der zweite ist das Crowding: Ein Nutzen, den jedes Haus aus dem Modell desselben Anbieters zieht, wird in den Kursen wegkonkurriert, wie Lo (2004) und Stein (2009) vorhersagen. Der dritte ist organisatorisch: Die messbaren Nutzen liegen in Betrieb, Reporting und Compliance, wo sie real, aber unspektakulär sind, während die versprochenen Nutzen im Alpha liegen, wo sie spektakulär, aber schwer fassbar sind.[43]

8.2 Scheinkorrelationen und Backtest-Überanpassung

Die zentrale methodische Gefahr der KI im Investieren ist älter als die KI. Granger und Newbold (1974) zeigten, dass Regressionen unverbundener trendbehafteter Reihen Scheinsignifikanz erzeugen; Lo und MacKinlay (1990) analysierten Verzerrungen durch Datenschnüffelei in Tests von Kapitalmarktmodellen; Sullivan, Timmermann und White (1999) sowie White (2000) zeigten, dass die scheinbare Rentabilität technischer Handelsregeln weitgehend verschwindet, sobald die Zahl der untersuchten Regeln berücksichtigt wird. Leinweber (2007) machte den Punkt unvergesslich, indem er den S&P 500 aus der Butterproduktion in Bangladesch „vorhersagte". Harvey, Liu und Zhu (2016) katalogisierten mehr als 300 veröffentlichte Faktoren und folgerten, dass ein neu vorgeschlagener Faktor angesichts der Zahl der Tests eine t-Statistik von mindestens drei, nicht zwei, überschreiten sollte. Flexible Modelle und billige Rechenleistung vervielfachen das Problem: Bailey et al. (2014) nannten die Praxis, den besten von vielen Backtests auszuwählen und als Entdeckung zu präsentieren, „Pseudomathematik und finanzielle Scharlatanerie" und zeigten, dass die erwartete maximale Sharpe-Ratio von N unverbundenen Strategien mit dem Logarithmus von N wächst, selbst wenn jede Strategie wertlos ist.

Abbildung 5 quantifiziert das Argument mit der Formel von Bailey et al. (2014) für den Erwartungswert des Maximums von N unabhängigen Sharpe-Ratios unter der Nullhypothese, dass die wahre Sharpe-Ratio jeder Variante null ist. Für T Jahre Tagesdaten ergibt sich die erwartete maximale annualisierte In-Sample-Sharpe-Ratio näherungsweise aus Gleichung (4),

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

wobei Z^{-1} die Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung, e die Eulersche Zahl und $\gamma_E \approx 0{,}5772$ die Euler-Mascheroni-Konstante bezeichnet; der Standardfehler der täglichen Sharpe-Ratio, $1/\sqrt{(252\,T)}$, und der Annualisierungsfaktor $\sqrt{252}$ ergeben zusammen $1/\sqrt{T}$.[44] Nach Gleichung (4) genügt das Testen von rund 45 Varianten

[43] Lo (2004) beschreibt die Verbreitung einer Innovation unter konkurrierenden Anlegern als jenen Mechanismus, durch den Überrenditen wegkonkurriert werden, in Analogie zum Wettbewerb der Arten in einem Ökosystem.

[44] Die einzelnen Rechenschritte sind in der begleitenden Excel-Datei dokumentiert.

mit fünf Jahren Tagesdaten, um eine erwartete beste annualisierte In-Sample-Sharpe-Ratio von eins zu erhalten; mit zwei Jahren Daten genügen rund acht Varianten, und tausend Varianten erzeugen eine erwartete beste Sharpe-Ratio über zwei (eigene Berechnungen). Ein genetischer Algorithmus oder eine Hyperparametersuche, die Tausende Konfigurationen bewertet, wird daher immer einen eindrucksvollen Backtest finden, und die Eindrücklichkeit trägt keinerlei Information über die künftige Leistung. Bailey und López de Prado (2014) schlugen die deflationierte Sharpe-Ratio vor, die die Zahl der Versuche korrigiert, und Arnott, Harvey und Markowitz (2019) schlugen ein Forschungsprotokoll für das Zeitalter des maschinellen Lernens vor, dessen erste Anforderung eine ökonomische Begründung jeder Hypothese ist.

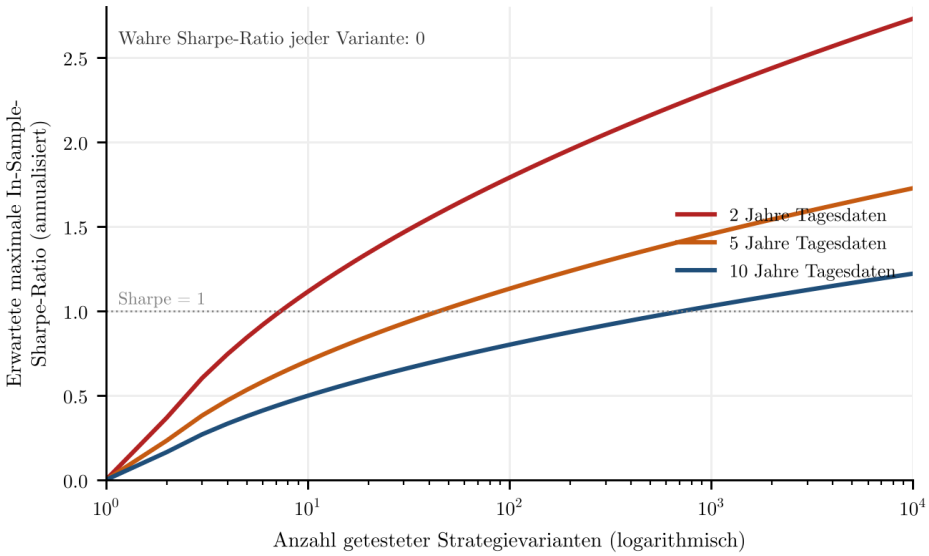


Abbildung 5: Erwartete maximale In-Sample-Sharpe-Ratio von N wertlosen Strategievarianten in Abhängigkeit von N
Quelle: Eigene Berechnungen auf Basis von Bailey et al. (2014).

Generative KI fügt dem Problem eine neue Form hinzu. Ein Sprachmodell, das gebeten wird, Zusammenhänge in einem Datensatz zu finden, wird sie finden und für jeden eine plausible ökonomische Erzählung liefern; die Erzählung entsteht aus dem Muster und nicht das Muster aus der Erzählung. Die Disziplin, die Breiman (2001) von der algorithmischen Kultur verlangte, die ehrliche Bewertung außerhalb der Stichprobe, und die Disziplin, die Arnott, Harvey und Markowitz (2019) von der Finanzkultur verlangten, eine ökonomische Begründung vor der Untersuchung der Daten, sind beide unverzichtbar, und die zweite kann nicht an jene Maschine delegiert werden, die das Muster erzeugt hat. Die Themenstellung dieser Arbeit hat es präzise formuliert: Reine Datenmodelle können falsche

Kausalitäten ableiten, weshalb eine ökonomische Fundierung und eine strenge Prüfung der Daten unerlässlich sind.[45]

8.3 Halluzination, Look-ahead-Bias und Erklärbarkeit

Drei Eigenschaften von Sprachmodellen erzeugen Modellrisiken, die klassische quantitative Modelle nicht hatten. Halluzination, die Erzeugung flüssiger Aussagen, die falsch oder unbelegt sind, ist Modellen inhärent, die wahrscheinlichen Text vorhersagen; Ji et al. (2023) sichten ihre Formen und Ursachen, und Bender et al. (2021) warnten früh, dass die Flüssigkeit solcher Modelle Leser dazu verleitet, Verständnis zu unterstellen, wo keines ist. Der Look-ahead-Bias, erörtert in Kapitel 4.3, entsteht, weil der Trainingskorpus eines Grundmodells die Ausgänge jener Ereignisse enthält, die ein Forscher vorhersagen möchte; die Evidenz von Glasserman und Lin (2023) sowie He et al. (2025) spricht dafür, dass der Bias in der Praxis gering ist, dass aber das Vorwissen eines Modells über bekannte Unternehmen seine Urteile verzerren kann. Die Erklärbarkeit ist schwächer als bei baumbasierten oder linearen Modellen: Verfahren wie die Shapley-Werte von Lundberg und Lee (2017) rechnen eine Prognose strukturierter Modelle ihren Eingaben zu, doch die Zurechnung eines erzeugten Absatzes zu den Milliarden Parametern, die ihn hervorgebracht haben, ist nicht sinnvoll, und Rudin (2019) argumentierte, dass Entscheidungen mit hohem Einsatz auf von sich aus interpretierbaren Modellen beruhen sollten statt auf nachträglichen Erklärungen von Black Boxes.[46]

Tabelle 7 fasst die Risiken mit ihren Mechanismen und den Vorkehrungen zusammen, die Praxis und Regulierung entwickelt haben. Keine der Vorkehrungen ist neu; alle sind die Grundsätze des Modellrisikomanagements aus der Aufsichtsleitlinie SR 11-7 der Federal Reserve und des Office of the Comptroller of the Currency (2011), angewandt auf eine neue Modellklasse: dokumentierter Zweck, unabhängige Validierung, laufende Überwachung und Grenzen für den Einsatz außerhalb des validierten Bereichs.

[45] Breiman (2001) beschrieb auch den Rashomon-Effekt: Viele verschiedene Modelle passen gleich gut zu denselben Daten, sodass die Wahl zwischen ihnen nicht allein aus den Daten getroffen werden kann.

[46] Shapley-Werte stammen aus der kooperativen Spieltheorie und verteilen eine Prognose nach dem durchschnittlichen Grenzbeitrag über alle möglichen Reihenfolgen auf die Eingabevariablen.

Tabelle 7: Risiken von KI und generativer KI im Asset Management, ihre Mechanismen und Vorkehrungen

Risiko	Mechanismus	Vorkehrung	Referenz
Backtest-Überanpassung	Auswahl des besten von vielen Versuchen; erwartete maximale Sharpe-Ratio steigt mit der Zahl der Versuche	Vorabfestlegung; deflationierte Sharpe-Ratio; zurückgehaltene Zeiträume; Protokoll aller Versuche	Bailey et al. (2014); Harvey, Liu und Zhu (2016)
Scheinkorrelation	Flexible Modelle passen sich dem Rauschen an; Erzählungen werden im Nachhinein erzeugt	Ökonomische Begründung vor dem Test; Validierung außerhalb der Stichprobe	Granger und Newbold (1974); Arnott, Harvey und Markowitz (2019)
Look-ahead-Bias	Trainingskorpus enthält Ausgänge des Testzeitraums	Trainingsende vor Testzeitraum; Maskierung von Entitäten; zeitpunktgenaue Modelle	Glasserman und Lin (2023); He et al. (2025)
Halluzination	Flüssige, aber falsche Aussagen; erfundene Zahlen und Quellen	Verankerung aller Zahlen in geprüften Daten; Quellenangabe; menschliche Prüfung	Ji et al. (2023)
Undurchsichtigkeit	Prognosen lassen sich nicht auf Ursachen zurückführen	Interpretierbare Modelle für Entscheidungen mit hohem Einsatz; Zurechnungsverfahren	Rudin (2019); Lundberg und Lee (2017)
Crowding	Viele Anleger nutzen dieselben Modelle und Signale	Vielfalt der Modelle; Kapazitätsgrenzen; Überwachung der Korrelation	Stein (2009); Khandani und Lo (2011)
Abhängigkeit von Dritten	Konzentration auf wenige Modell- und Cloud-Anbieter	Vertragliche und technische Ersatzlösungen; IKT-Risikomanagement nach DORA	Financial Stability Board (2024); Europäische Union (2022)
Regimewechsel	Auf ein Regime trainierte Modelle versagen in einem anderen	Stresstests; menschliches Eingreifen; ökonomische Vorannahmen	Israel, Kelly und Moskowitz (2020)

Quelle: Eigene Darstellung.

8.4 Crowding, Homogenisierung und Systemrisiko

Die Sorgen der Aufsicht reichen über das einzelne Haus hinaus. Das Financial Stability Board (2017) erkannte früh das Potenzial neuer Formen der Vernetzung und des Aufstiegs weniger dominanter Anbieter; sein Bericht von 2024 ergänzte die Verwundbarkeiten durch Abhängigkeiten von Dritten und die Konzentration von Dienstleistern, durch erhöhte Marktkorrelationen infolge ähnlicher Modelle sowie durch Cyber- und Modellrisiken (Financial Stability Board, 2024). Der Internationale Währungsfonds (2024) widmete der KI an den Kapitalmärkten ein Kapitel seines Global Financial Stability Report und kam zu dem Schluss, dass KI die Märkte effizienter, aber auch schneller und unter Stress volatiler machen könnte und dass die Undurchsichtigkeit der Modelle und die Abhängigkeit von wenigen Anbietern eine engere Überwachung rechtfertigen. Die Europäische Zentralbank (2024) gelangte für den Euroraum zu einer ähnlichen Einschätzung, und

IOSCO (2021) formulierte sechs Maßnahmen für Marktintermediäre und Asset Manager, von der Verantwortung des Senior Managements bis zur Kontrolle ausgelagerter KI.[47]

Die wissenschaftliche Evidenz ist noch jung, weist aber in dieselbe Richtung. Dou, Goldstein und Ji (2025) zeigten, dass Händler auf Basis des verstärkenden Lernens, von denen jeder seinen eigenen Gewinn maximiert, ohne jede Kommunikation zu kollusivem Verhalten konvergieren, das den Informationsgehalt der Kurse verringert, und übertrugen damit die Ergebnisse von Calvano et al. (2020) zur algorithmischen Preissetzung auf die Finanzmärkte. Kirilenko et al. (2017) dokumentierten, wie das Zusammenwirken algorithmischer Händler den Flash Crash verstärkte.[48] Wenn generative KI Tausende Manager dazu bringt, dieselben Dokumente mit denselben Modellen zu lesen und im selben Moment zu denselben Schlüssen zu kommen, sind der Effizienzgewinn aus Kapitel 3 und die hier beschriebene Zerbrechlichkeit zwei Seiten desselben Phänomens.

8.5 Der regulatorische Rahmen

Der regulatorische Rahmen für KI im europäischen Asset Management hat vier Ebenen, die Tabelle 8 zusammenfasst. Die erste ist die KI-Verordnung, Verordnung (EU) 2024/1689, in Kraft seit 1. August 2024, die einem risikobasierten Ansatz folgt: eine kurze Liste verbotener Praktiken, ein Katalog von Hochrisiko-Anwendungen in Anhang III, Transparenzpflichten für Systeme, die mit Menschen interagieren oder Inhalte erzeugen, und Pflichten für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck (Europäische Union, 2024). Für Asset Manager ist entscheidend, dass Portfolioverwaltung und Anlageberatung nicht als Hochrisiko-Anwendungen gelistet sind; die für Finanzinstitute relevanten Hochrisiko-Tatbestände sind die Bewertung der Kreditwürdigkeit natürlicher Personen sowie Risikobewertung und Preisbildung in der Lebens- und Krankenversicherung. Die meisten Anlageanwendungen fallen daher unter die Transparenz- und KI-Kompetenzpflichten und unter die allgemeinen Pflichten des Finanzaufsichtsrechts und nicht unter das Hochrisiko-Regime. Der Digital Omnibus zur KI, Verordnung (EU) 2026/1744, veröffentlicht am 24. Juli 2026, verschob die Anwendung der Pflichten nach Anhang III vom 2. August 2026 auf den 2. Dezember 2027 und jener nach Anhang I auf den 2. August 2028, milderte die KI-Kompetenzpflicht des Artikels 4 zu einer Pflicht, unterstützende Maßnahmen zu ergreifen, und führte Erleichterungen für

[47] Die Maßnahmen von IOSCO (2021) umfassen Governance und Verantwortung des Senior Managements, Tests und Überwachung, Kompetenz des Personals, Kontrolle von Drittanbietern, Offenlegung gegenüber Kunden und Aufsicht sowie Datenqualität.

[48] Kirilenko et al. (2017) kommen zu dem Schluss, dass Hochfrequenzhändler den Flash Crash vom 6. Mai 2010 nicht ausgelöst, ihn aber verstärkt haben, indem sie Liquidität entzogen und in Richtung der Kursbewegung handelten.

kleine Unternehmen mittlerer Kapitalisierung ein (Europäische Union, 2026). Abbildung 6 zeigt den daraus folgenden Zeitplan.[49]

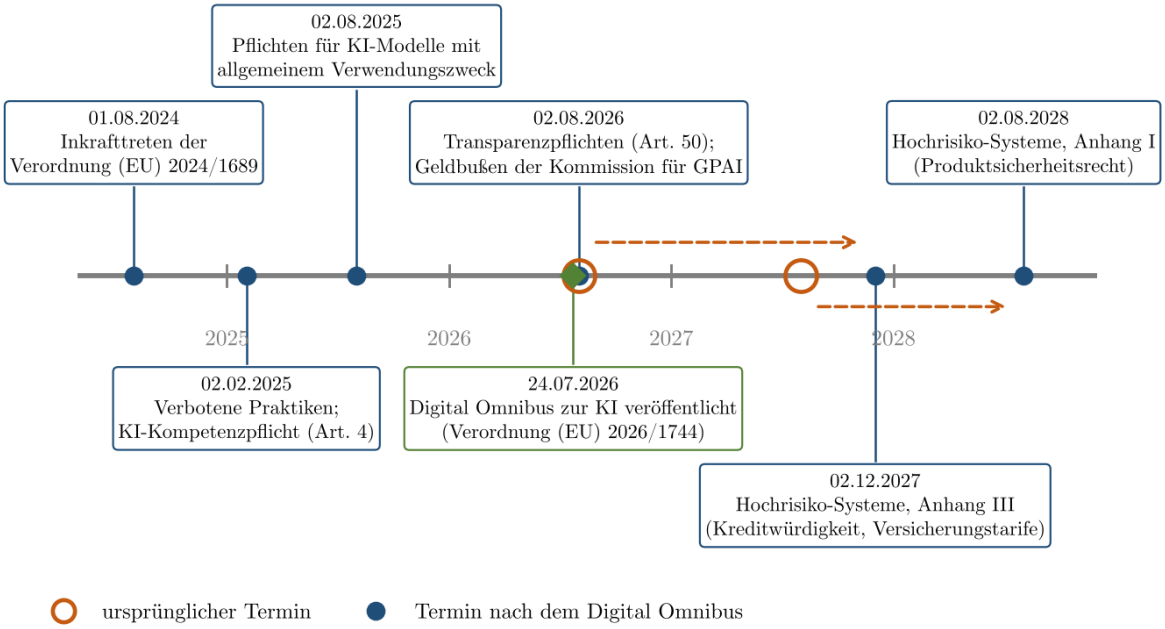


Abbildung 6: Anwendungstermine der EU-KI-Verordnung nach dem Digital Omnibus vom Juli 2026
Quelle: Eigene Darstellung auf Basis von Europäische Union (2024) und Europäische Union (2026).

Die zweite Ebene ist das sektorale Finanzaufsichtsrecht. MiFID II (Europäische Union, 2014) und ihre delegierten Rechtsakte regeln Eignung, bestmögliche Ausführung, Interessenkonflikte, Aufzeichnungspflichten und die organisatorischen Anforderungen, die für jede Technologie gelten; ESMA (2024) hat ausdrücklich klargestellt, dass sie für KI gelten. Die dritte Ebene ist die operationelle Widerstandsfähigkeit: Der Digital Operational Resilience Act, Verordnung (EU) 2022/2554, anwendbar seit 17. Januar 2025, verpflichtet Finanzunternehmen, die Risiken von IKT-Drittdienstleistern zu managen, was die Anbieter von Grundmodellen und der Cloud-Infrastruktur einschließt, auf der sie laufen (Europäische Union, 2022). Die vierte Ebene ist der Datenschutz: Die Datenschutz-Grundverordnung beschränkt die Verarbeitung personenbezogener Daten beim Training und bei der Anweisung von Modellen und in Artikel 22 ausschließlich maschinell getroffene Entscheidungen mit rechtlicher Wirkung gegenüber Einzelpersonen (Europäische Union, 2016). Für ein österreichisches Wertpapierunternehmen unter Aufsicht der Finanzmarktaufsicht gelten alle vier Ebenen zugleich, und die KI-Kompetenz des

[49] Anhang III nennt auch beschäftigungsbezogene Anwendungen wie Personalauswahl und Leistungsbeurteilung, die Asset Manager als Arbeitgeber betreffen, nicht als Anleger.

Personals, das Verzeichnis der eingesetzten KI-Systeme und die Dokumentation der menschlichen Aufsicht sind die praktischen Ausgangspunkte der Compliance.[50]

Tabelle 8: Regulatorischer Rahmen für KI im europäischen Asset Management

Rechtsakt	Anwendungsbereich	Bedeutung für Asset Manager	Wesentliche Termine
KI-Verordnung, Verordnung (EU) 2024/1689, geändert durch Verordnung (EU) 2026/1744	Risikobasierte Regeln für KI-Systeme und KI-Modelle mit allgemeinem Verwendungszweck	Portfolioverwaltung kein Hochrisiko; Transparenz- und KI-Kompetenzpflichten; Kreditwürdigkeit und Versicherungstarife Hochrisiko	In Kraft 1.8.2024; GPAI 2.8.2025; Anhang III 2.12.2027; Anhang I 2.8.2028
MiFID II, Richtlinie 2014/65/EU	Wertpapierdienstleistungen und Anlagetätigkeiten	Eignung, bestes Interesse, organisatorische Anforderungen, Aufzeichnungspflichten gelten für KI-gestützte Dienstleistungen	Anwendbar seit 3.1.2018
Öffentliche Erklärung der ESMA zur KI (2024)	Wertpapierdienstleistungen für Kleinanleger	Verantwortung des Leitungsorgans; Tests und Überwachung; Schulung des Personals; Kundeninformation	30.5.2024
DORA, Verordnung (EU) 2022/2554	Digitale operationelle Widerstandsfähigkeit von Finanzunternehmen	IKT-Drittparteienrisikomanagement umfasst Modell- und Cloud-Anbieter	Anwendbar seit 17.1.2025
DSGVO, Verordnung (EU) 2016/679	Verarbeitung personenbezogener Daten	Training und Anweisung mit Kundendaten; Art. 22 ausschließlich maschinelle Entscheidungen	Anwendbar seit 25.5.2018
SR 11-7 (Federal Reserve und Office of the Comptroller of the Currency, 2011)	Modellrisikomanagement (US-Aufsichtsleitlinie)	Faktischer internationaler Standard für Validierung und Governance von Modellen	April 2011

Quelle: Eigene Darstellung auf Basis der zitierten Rechtsakte und Europäische Union (2026).

8.6 Governance: Autonomiegrade und menschliche Aufsicht

Die regulatorischen Ebenen laufen in einer einzigen Governance-Frage zusammen: Wie viel Autonomie darf ein KI-System in einer bestimmten Anwendung haben, und welche Form muss die menschliche Aufsicht annehmen? Abbildung 7 schlägt eine einfache Antwort vor. Der vertretbare Grad an Autonomie sinkt, wenn die Folgen eines Fehlers für den Kunden steigen. Das Zusammenfassen von Dokumenten und das Entwerfen von Texten, die eine Person prüft, liegen im Bereich der Assistenz, wo die Kosten eines Fehlers gering und seine Entdeckung leicht sind. Die Überwachung von Richtlinien und das Erzeugen von Warnhinweisen liegen in der Mitte: Das System handelt, aber eine Person entscheidet. Autonome Portfolioentscheidungen für einen Kunden liegen in jenem Bereich, in dem eine menschliche Entscheidung zwingend ist, nicht weil Maschinen notwendigerweise schlechter wären, sondern weil die Verantwortung nach MiFID II und die treuhänderische Beziehung

[50] Die österreichische Finanzmarktaufsicht beaufsichtigt Wertpapierunternehmen nach dem Wertpapieraufsichtsgesetz 2018 (WAG 2018), das MiFID II umsetzt; die hier beschriebenen Pflichten gelten für österreichische Unternehmen daher ohne weitere nationale KI-spezifische Regeln.

nicht auf sie übertragen werden können und weil die Systemrisiken aus Kapitel 8.4 mit dem Anteil der von ähnlichen Modellen getroffenen Entscheidungen wachsen.[51]

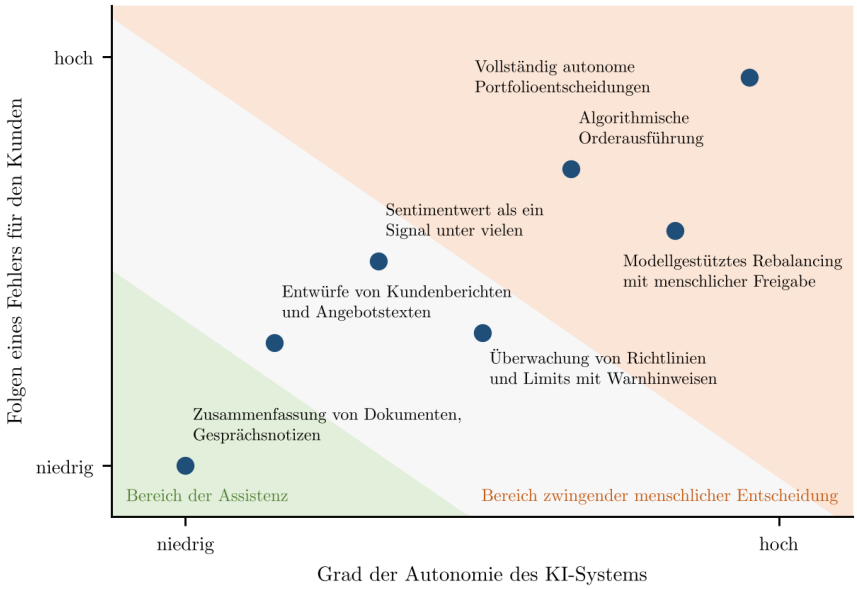


Abbildung 7: Vertretbarer Autonomiegrad von KI-Systemen im Asset Management nach den Folgen eines Fehlers
Anmerkungen: Vom Verfasser vorgeschlagener konzeptioneller Rahmen; die Positionen der Anwendungen sind qualitative Einschätzungen, keine Messungen.
Quelle: Eigene Darstellung.

Operativ bedeutet dies dreierlei. Jedes eingesetzte KI-System wird mit seinem Zweck, seinen Daten, seinem Anbieter und seinem validierten Bereich in einem Verzeichnis geführt. Jede Ausgabe, die eine Entscheidung beeinflusst, trägt ihre Quellen und wird von einer Person geprüft, deren Kompetenz dafür dokumentiert ist, was der Kern der KI-Kompetenzpflicht ist. Und jede Entscheidung, die einen Kunden betrifft, wird von einer Person getroffen und als von ihr getroffen festgehalten. Innerhalb dieser Grenzen kann der Grad an Autonomie mit der Erfahrung steigen; außerhalb dieser Grenzen legen die Erfahrungen vom August 2007 und vom Mai 2010 nahe, dass die Kosten der Autonomie von anderen getragen werden.

9 Fazit und Ausblick

Diese Arbeit hat sich vorgenommen zu untersuchen, wie generative KI im Vergleich zu traditionellen quantitativen Modellen die Informationseffizienz der Aktienmärkte und die Praxis des Asset Managements verändert. Das theoretische Argument ist im Rahmen von Grossman und Stiglitz (1980) einfach: Klassische Prognosemodelle erhöhen die Präzision

[51] Die KI-Verordnung verlangt menschliche Aufsicht (Art. 14) nur für Hochrisiko-Systeme; für andere Systeme folgt die Anforderung aus den organisatorischen Regeln der MiFID II und aus der zivilrechtlichen Haftung.

privater Signale und damit den Anreiz, Information zu beschaffen, während generative Modelle die Kosten der Verarbeitung öffentlicher Information, insbesondere von Text, senken und damit jene Reibungen, die Blankespoor, deHaan und Marinovic (2020) als Verarbeitungskosten von Offenlegungen beschrieben haben. Die Vorhersage erster Ordnung lautet, dass generative KI den qualitativen Inhalt von Earnings Calls und Berichten schneller und vollständiger in die Blue-Chip-Kurse eingehen lässt und die Drift verkürzt, die Bernard und Thomas (1989), Price et al. (2012) und Cohen, Malloy und Nguyen (2020) dokumentierten, während ihre Wirkung auf den langfristigen Informationsgehalt der Kurse unbestimmt ist.[52]

Die empirische Literatur stützt die Prämisse. Von Loughran und McDonald (2011) über Huang, Wang und Yang (2023) bis zu Lopez-Lira und Tang (2023) und Chen, Kelly und Xiu (2022) hat jede Generation textbasierter Verfahren mehr renditerelevante Information aus denselben Dokumenten gewonnen, und die generative Generation tut dies mit beispielloser Breite und Geschwindigkeit. Dieselbe Literatur mahnt zugleich zur Sorgfalt: Eine prominente Studie wurde von ihren Autoren zurückgezogen, nachdem Unstimmigkeiten in den Daten entdeckt worden waren (Kim, Muhn und Nikolaev, 2024), und die Tests von Glasserman und Lin (2023) zeigen, dass das Vorwissen eines Modells über Unternehmen seine Urteile verzerren kann. Das Forschungsdesign in Kapitel 5 macht aus diesen Lehren eine konkrete Studie. Wir erwarten, als Hypothese und nicht als Ergebnis, eine messbare und abnehmende Drift nach LLM-basierten Sentiment-Überraschungen bei Blue-Chip-Aktien, was ein unmittelbarer Beleg für steigende Informationseffizienz wäre.

Für die Praxis sind die Schlussfolgerungen vorsichtiger als der öffentliche Diskurs. Im Fondsmanagement ist generative KI zu einer unentbehrlichen Lesemaschine geworden, die die Abdeckung erweitert und strukturierte Variablen in etablierte Signalmodelle einspeist; eine Generation überlegener autonomer Fonds hat sie nicht hervorgebracht, und die größten Manager setzen sie durchgängig ein, belassen die Portfolioentscheidung aber beim Menschen. In Einzelmandaten liegen ihre Vorteile vor allem in der personalisierten Kommunikation, sofern jede Zahl in geprüften Daten verankert ist. Die Studie von KPMG (2026) ist ein notwendiges Korrektiv: Nur rund eines von acht Häusern hat einen messbaren Ertrag erzielt.[53] Die Statistik von Bailey et al. (2014) ist ein zweites Korrektiv: Nur eine vorab formulierte ökonomische Begründung und ein ehrlicher Test außerhalb der Stichprobe trennen Erkenntnis von Zufall.

Der Ausblick hat drei Dimensionen. Technologisch wird sich der Übergang von Assistenten zu Agenten fortsetzen, und Modelle mit offenen Gewichten und dokumentierten Trainingskorpora werden belastbare Forschungsdesigns erleichtern.

[52] Fama (1970) fasste Effizienz als Aussage über Geschwindigkeit und Vollständigkeit der Kursanpassung, weshalb die Drift und nicht die Ankündigungsrendite der natürliche Gegenstand des Tests ist.

[53] Eines von acht entspricht den 12 Prozent der Befragten bei KPMG (2026), die skalierte KI mit messbarem Return on Investment melden.

Ökonomisch wird der Vorteil generativer KI vom Lesen selbst, das alle tun werden, zu den proprietären Daten wandern, die gelesen werden, und zum Urteil, das darauf angewandt wird; die stärksten Effizienzgewinne sollten bei Nebenwerten mit dünner Abdeckung auftreten, wo die nächste empirische Arbeit ansetzen sollte. Regulatorisch hat sich der europäische Rahmen auf eine Position verständigt, die diese Arbeit teilt: KI darf lesen, vorschlagen und überwachen, während ein Mensch entscheidet und für die Entscheidung einsteht. Generative KI stärkt ökonomisch fundiertes Urteil; sie ersetzt es nicht, und wer den Unterschied vergisst, wird ihn in jener Phase entdecken, für die seine Modelle keine Trainingsdaten hatten.

Literaturverzeichnis

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 und 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. Juli 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. Mai 2023.

CNBC (2025). Goldman Sachs rolls out an AI assistant for its employees as artificial intelligence sweeps Wall Street (H. Son). CNBC, 21. Januar 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. Mai 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. Pressemitteilung, 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. Mai 2024.

Europäische Union (2014). Richtlinie 2014/65/EU des Europäischen Parlaments und des Rates vom 15. Mai 2014 über Märkte für Finanzinstrumente (MiFID II). *Amtsblatt der Europäischen Union*, L 173, 349–496.

Europäische Union (2016). Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 (Datenschutz-Grundverordnung). *Amtsblatt der Europäischen Union*, L 119, 1–88.

Europäische Union (2017). Delegierte Verordnung (EU) 2017/565 der Kommission vom 25. April 2016 zur Ergänzung der Richtlinie 2014/65/EU in Bezug auf die organisatorischen Anforderungen an Wertpapierfirmen und die Bedingungen für die Ausübung ihrer Tätigkeit. *Amtsblatt der Europäischen Union*, L 87, 31. März 2017, 1–83.

Europäische Union (2022). Verordnung (EU) 2022/2554 des Europäischen Parlaments und des Rates vom 14. Dezember 2022 über die digitale operationale Resilienz im Finanzsektor (DORA). *Amtsblatt der Europäischen Union*, L 333, 1–79.

Europäische Union (2024). Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche

Intelligenz (Verordnung über künstliche Intelligenz). *Amtsblatt der Europäischen Union*, L, 12. Juli 2024.

Europäische Union (2026). Verordnung (EU) 2026/1744 des Europäischen Parlaments und des Rates zur Änderung der Verordnung (EU) 2024/1689 (Digital Omnibus zur KI). *Amtsblatt der Europäischen Union*, 24. Juli 2026.

Europäische Zentralbank (2024). The rise of artificial intelligence: Benefits and risks for financial stability. *Financial Stability Review*, Mai 2024, Sonderbeitrag (15. Mai 2024).

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 und 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* (2. Aufl.). 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.

Internationaler Währungsfonds (2024). Advances in artificial intelligence: Implications for capital market activities. *Global Financial Stability Report*, Oktober 2024, Kapitel 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), 2. Quartal 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 (von den Autoren am 20. Februar 2025 zurückgezogen).

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 (Hrsg.), *Handbook of Empirical Corporate Finance* (Vol. 1, S. 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 (abgerufen am 29. Oktober 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 (überarbeitete Fassung Oktober 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. Nachdruck 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. Juni 2023.

Mercer (2024). AI integration in investment management: 2024 global manager survey. Mercer Investments, März 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. Oktober 2025.

Morgan Stanley (2024). Morgan Stanley Wealth Management announces latest game-changing addition to suite of GenAI tools. Pressemitteilung, 26. Juni 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. Fallstudie, OpenAI, 6. März 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. Juli 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). Mensch oder Maschine? Wo Künstliche Intelligenz in der Vermögensanlage Mehrwert liefert und wo menschliches Verhalten überlegen bleibt – Allokation und Titelselektion bei privaten und professionellen Anlegern. 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. Juli 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 mit Kevin Endler. DAS INVESTMENT, 16. Mai 2025.

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

Union Investment (2024). Union Investment startet Vermögensverwaltung mit Künstlicher Intelligenz. Pressemitteilung, 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.