universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Mensch vs. Künstliche Intelligenz: Evidenz aus der Verarbeitung von Finanzberichten

verfasst von | submitted by

Aleksa Ljubicic BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 915
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Betriebswirtschaft
Betreut von | Supervisor:	Assoz. Prof. Daniel Garcia PhD

Abstract

Die vorliegende Arbeit untersucht, ob ChatGPT auf Basis von Quartalsberichten (Form 10-Q) Handlungsempfehlungen für Aktien generieren kann, die mit jenen professioneller Finanzanalysten vergleichbar sind. Dazu werden für 50 S&P-500-Unternehmen die Empfehlungen von ChatGPT-5.4 Thinking mit dem Analystenkonsens verglichen und anhand der realisierten Aktienkursentwicklung in vier Prognosehorizonten (10, 30, 60 und 120 Handelstage) bewertet. Die Ergebnisse zeigen, dass GPT auf kurzfristigen und mittelfristigen Horizonten statistisch nicht signifikant schlechter als Analysten abschneidet. Lediglich im längsten Horizont von 120 Handelstagen erweisen sich Analysten als schwach signifikant treffsicherer. Eine zusätzliche Rollenzuweisung an GPT als Aktienanalyst verbessert die Performance nicht. Die Befunde legen nahe, dass ChatGPT als unterstützendes Analyseinstrument im Finanzbereich Potenzial besitzt, bei längeren Horizonten jedoch strukturellen Limitationen unterliegen könnte.

Inhaltsverzeichnis

1 Einleitung

Die zunehmende Verfügbarkeit leistungsfähiger Large Language Models (LLMs) hat in den vergangenen Jahren zahlreiche Bereiche der Wissensarbeit verändert. Systeme wie ChatGPT können nicht nur allgemeine Sprachaufgaben wie Zusammenfassungen oder Klassifikationen bewältigen, sondern werden zunehmend auch in spezialisierten Domänen wie dem Finanz- und Rechnungswesen eingesetzt (Dong et al., 2024). Gerade hier ist diese Entwicklung von besonderem Interesse: Finanzentscheidungen erfordern die Integration großer Mengen strukturierter und unstrukturierter Information unter Zeit- und Unsicherheitsdruck, genau jene Bedingungen, unter denen LLMs ihr Potenzial entfalten können (Aleksandrova et al., 2023).

Unternehmensberichte wie Quartals- und Jahresabschlüsse enthalten eine Vielzahl entscheidungsrelevanter Informationen. Einerseits vermitteln sie numerische Kennzahlen zu Umsatz, Gewinn oder Verschuldung, andererseits enthalten sie narrative Einschätzungen des Managements zu Risiken und Zukunftsaussichten. Jahresberichte werden immer umfangreicher und die Menge unstrukturierter Finanztexte wächst so stark, dass Fachleute oft nicht mehr mithalten können, diese Daten zeitnah zu verarbeiten (Hao, 2025). Genau an dieser Stelle entsteht die Frage, ob moderne Sprachmodelle in der Lage sind, Finanzberichte so zu verarbeiten, dass daraus zuverlässige und potenziell marktrelevante Handlungsempfehlungen resultieren.

Die Forschung zu LLMs im Finanzkontext hat in den letzten Jahren deutlich zugenommen. Es wurden Studien veröffentlicht, in denen LLMs zur Sentimentanalyse, Textgenerierung, Klassifikation und Prognose von Finanzkennzahlen eingesetzt werden (Dong et al., 2024). Lopez-Lira und Tang (2025) zeigen etwa, dass ChatGPT aus Nachrichtenüberschriften mit einer hohen Trefferquote die kurzfristige Kursrichtung vorhersagt. Li et al. (2024) belegen, dass GPT-Prognosen zu Finanzkennzahlen im Durchschnitt näher an den realisierten Werten liegen als jene professioneller Analysten und dabei konservativer ausfallen, was dem bekannten Optimismus-Bias der Analysten entgegenwirkt. Bilinski (2024) demonstriert, dass von ChatGPT generierte Sentiment- und Komplexitätsscores aus Jahresberichten wirtschaftlich sinnvolle Informationen über Marktpreise und Zukunftserträge enthalten. Gleichzeitig verweist die Literatur auf zentrale Einschränkungen. Numerische Schwächen bei komplexen und langen Finanzdokumenten (Zhao et al., 2024), potenzielle Halluzinationen sowie die Sensitivität gegenüber dem Prompting-Design (Barman et al., 2024).

Trotz dieser wachsenden Forschungslage fehlt bislang ein direkter Vergleich zwischen LLM-generierten Handlungsempfehlungen und professionellen Analystenurteilen auf Basis realer Quartalsberichte, insbesondere mit dem leistungsfähigsten aktuell öffentlich verfügbaren Modell GPT-5.4 Thinking. Aus diesem Forschungsdesiderat ergeben sich die zwei zentralen Hypothesen dieser Arbeit. **Hypothese 1** lautet, dass sich die auf Basis von Quartalsreports generierten Handlungsempfehlungsratings der Analysten von jenen von ChatGPT unterscheiden. **Hypothese 2** lautet, dass ChatGPT im Hinblick auf die spätere Aktienkursentwicklung weniger zuverlässige Handlungsempfehlungen abgibt als Analysten und zwar kurzfristig, mittelfristig sowie langfristig.

Zur empirischen Prüfung dieser Hypothesen werden 50 Unternehmen des S&P 500 untersucht. Für jedes Unternehmen wird der erste Quartalsbericht nach dem Trainings-Cut-Off des Modells an ChatGPT-5.4 Thinking übermittelt, das daraus eine Handlungsempfehlung auf einer fünfstufigen Ratingskala (Strong Buy bis Strong Sell) ableitet. Die Wahl des Thinking-Modus erscheint für diese Aufgabe am geeignetsten, da die Analyse eines vollständigen 10-Q-Berichts nicht nur eine schnelle Antwort, sondern die strukturierte Verarbeitung umfangreicher quantitativer und qualitativer Informationen erfordert (OpenAI API, o. J.). Die GPT-Empfehlungen werden mit dem zur Veröffentlichung der jeweiligen Quartalsreports zeitnahen Analystenkonsens aus LSEG Workspace verglichen, der die aggregierten Bewertungen professioneller Analystenhäuser widerspiegelt. Anschließend werden beide Seiten anhand der realisierten Kursverläufe in vier Prognosefenstern bewertet: 10, 30, 60 und 120 Handelstage nach einem angenommenen Empfehlungs-Berichtsstichtag der Analysten. Ergänzend werden Robustheitschecks durchgeführt, unter Einbeziehung von Hold-Ratings, unter Berücksichtigung der unterschiedlich möglichen Reaktionszeit der Analysten sowie mit einem zusätzlichen GPT-Durchlauf, in dem dem Modell explizit die Rolle eines professionellen Aktienanalysten zugewiesen wird.

Die Arbeit leistet damit einen Beitrag zu zwei Forschungssträngen. Einerseits erweitert sie die Literatur zur finanziellen Informationsverarbeitung durch LLMs um einen direkten Vergleich zwischen ChatGPT und professionellen Analysten auf Basis standardisierter Quartalsberichte. Andererseits liefert sie Evidenz dazu, auf welchen Prognosehorizonten LLMs bereits konkurrenzfähig erscheinen und wo strukturelle Grenzen bestehen, eine Frage mit unmittelbarer Relevanz für die praktische Anwendung von KI-Systemen in der Finanzanalyse.

Der weitere Aufbau der Arbeit gliedert sich wie folgt: Zunächst werden die theoretischen Grundlagen zu künstlicher Intelligenz, Large Language Models und der Transformer-Architektur dargestellt. Darauf aufbauend wird der Forschungsstand zur Verarbeitung numerischer und narrativer Finanzinformationen durch LLMs sowie zu deren Einsatz in der Finanzanalyse aufgearbeitet. Anschließend werden Datenbasis, Untersuchungsdesign und methodisches Vorgehen erläutert, gefolgt von den Ergebnissen einschließlich der Robustheitschecks und der Regressionsanalysen. Abschließend werden die Befunde diskutiert, Limitationen aufgezeigt und Implikationen für zukünftige Forschung abgeleitet.

2 Künstliche Intelligenz, Sprachverarbeitung und Large Language Modelle

Künstliche Intelligenz (KI) bezeichnet Systeme, die Aufgaben ausführen, die traditionell menschliche Intelligenz erfordern, etwa Mustererkennung, Lernen aus Daten oder Sprachverstehen. Für diese Arbeit ist ein funktionales Verständnis ausreichend: KI umfasst Verfahren, die aus Daten Beziehungen ableiten und Probleme bearbeiten, die sich nicht vollständig über starre Regeln programmieren lassen (König et al., 2022). In der Praxis dominieren schwache KI-Systeme, die auf klar abgegrenzte Aufgaben spezialisiert sind, etwa in der Bilderkennung, bei Empfehlungssystemen oder in der Sprachverarbeitung (König et al., 2022).

Maschinelles Lernen bildet den zentralen technischen Baustein moderner KI. Modelle werden anhand historischer Daten trainiert, um statistische Muster zu erkennen, statt ausschließlich expliziten Regeln zu folgen (Welsch et al., 2018). Besonders relevant für diese Arbeit ist die Verarbeitung natürlicher Sprache (Natural Language Processing, NLP). Ziel von NLP ist es, Texte so zu repräsentieren, dass Computersysteme Bedeutungen, Beziehungen und Handlungsimplikationen ableiten können, etwa für Klassifikation, Zusammenfassung, maschinelle Übersetzung oder Sentimentanalyse (X. Cheng et al., 2025, S. 1 ff.). Dies ist im Finanzkontext zentral, da ein großer Teil entscheidungsrelevanter Information in Textform vorliegt, etwa in Geschäfts- und Quartalsberichten, Ad-hoc-Meldungen oder Analystenkommentaren.

2.1 Large Language Modelle

Large Language Models (LLMs) sind große neuronale Sprachmodelle, die im Rahmen eines umfangreichen Pre-Trainings lernen, Token-Sequenzen in Texten vorherzusagen. Sie modellieren die Wahrscheinlichkeit, welches Wort oder Token in einem gegebenen Kontext als Nächstes auftreten sollte, unterscheiden sich von früheren Sprachmodellen jedoch durch ihre Größe, den Umfang der Trainingsdaten und die Fähigkeit, lange Kontexte zu berücksichtigen (Jurafsky & Martin, 2024, S. 145 ff.).

Ein LLM erhält einen Textkontext und berechnet eine Wahrscheinlichkeitsverteilung über mögliche nächste Tokens, anschließend wird ein Token ausgewählt und an die Sequenz angehängt. Die Eingabe des Nutzers wird als Prompt bezeichnet und kann Anweisungen, Fragen oder strukturierte Aufgaben umfassen. Unterschiedliche Prompting-Strategien, etwa Zero-Shot-Prompts ohne Beispiele, Few-Shot-Prompts mit Beispielaufgaben oder Chain-of-Thought-Prompts mit expliziter Zwischenschritt-Anforderung beeinflussen Genauigkeit und Nachvollziehbarkeit der Antworten (Jurafsky & Martin, 2024, S. 150 ff.; J. Li et al., 2025). Für finanzielle Anwendungen ist dies relevant, weil die Promptgestaltung mitentscheidet, wie zuverlässig LLM-basierte Analysen ausfallen.

Die Auswahl des nächsten Tokens (Decoding) kann deterministisch oder stochastisch erfolgen. Greedy Decoding wählt stets das wahrscheinlichste Token und erzeugt stabile, aber oft repetitive Ausgaben, während Verfahren wie Temperature-, Top-k- oder Top-p-Sampling auch weniger wahrscheinliche Tokens zulassen und so die Varianz der Antworten erhöhen (Jurafsky & Martin, 2024, S. 154 ff.). Dadurch können identische Prompts in verschiedenen Durchläufen leicht unterschiedliche Antworten erzeugen.
Das breite Einsatzspektrum von LLMs basiert auf ihrem Pre-Training auf großen Textkorpora wie Webtexten, Wikipedia oder Büchern. In diesem Schritt erlernen die Modelle Sprachmuster, Weltwissen und fachliche Begriffe (Jurafsky & Martin, 2024, S. 160 ff.).

2.2 Transformer-Architektur

Die von Vaswani et al. (2017) vorgeschlagene Transformer-Architektur bildet die technische Basis moderner LLMs. Anders als andere Modelle verarbeitet der Transformer Eingabesequenzen mithilfe eines Attention-Mechanismus, durch den jedes Token direkt mit allen anderen Tokens in Beziehung gesetzt wird. Self-Attention erlaubt es, für jedes Token zu

gewichten, welche Teile der Sequenz für dessen Interpretation besonders wichtig sind, sodass auch weit entfernte Kontextbezüge effizient berücksichtigt werden (Vaswani et al., 2017). Da Tokens parallel verarbeitet werden können, ist das Training deutlich effizienter als bei anderen Modell-Architekturen.

Auf Basis dieser Transfomer-Architektur entstanden Modelle wie BERT und die GPT-Reihe. Die GPT-Modelle („Generative Pre-Trained Transformer“) kombinieren großskaliges Pre-Training mit der Transformer-Architektur und können ein breites Spektrum an Aufgaben bewältigen, von Klassifikation über Fragenbeantwortung bis hin zur kohärenten Textgenerierung (Islam et al., 2024). Durch zusätzliches Training mit menschlichem Feedback wurden InstructGPT und darauf aufbauend ChatGPT entwickelt, die besser auf Instruktionen reagieren und dialogfähig sind (Ouyang et al., 2022).

2.2.1 ChatGPT

ChatGPT ist eine dialogorientierte Variante der GPT-Reihe, die durch nutzerorientiertes Feintuning auf hilfreiche, inhaltlich sinnvolle und höfliche Antworten ausgerichtet wurde (OpenAI, 2022; Ouyang et al., 2022). Gleichzeitig weisen die Entwickler darauf hin, dass ChatGPT überzeugende, aber inhaltlich falsche Argumentationen erzeugen und bei unklaren Anfragen spekulativ agieren kann (OpenAI, 2022).
Aus der GPT-Reihe gingen die Modelle GPT-1 bis GPT-5 hervor, die mit sehr großen Textmengen überwiegend im Rahmen unüberwachten Lernens trainiert wurden (Islam et al., 2024; Radford et al., 2019). Mit zunehmender Modellgröße und Datenbasis konnten GPT-2 und GPT-3 bereits ohne aufgabenspezifisches Fine-Tuning vielfältige Aufgaben lösen. Auf GPT-3 folgte 2023 GPT-4. Zum Zeitpunkt dieser Arbeit ist GPT-5.4 das aktuelle Modell, das auch in dieser Studie verwendet wird (OpenAI, o. J.-a). In ChatGPT wird zwischen Nutzungsmodi wie Instant und Thinking unterschieden. Instant ist auf schnelle Alltagsanfragen ausgelegt, während Thinking für komplexere Aufgaben mit vertieftem Reasoning vorgesehen ist (OpenAI, o. J.-a). Für die vorliegende Untersuchung kommt explizit GPT-5.4 Thinking zum Einsatz, da die Analyse vollständiger Quartalsberichte lange Kontexte und die Integration numerischer sowie narrativer Informationen erfordert.

Zusammenfassend, ist die Transformer-Architektur damit nicht nur technisch, sondern auch konzeptionell relevant. Finanzberichte enthalten eng miteinander verwobene Kennzahlen,

narrative Erläuterungen, Risikohinweise und Managementeinschätzungen. Ein Modell, das lange Kontexte verarbeiten und Beziehungen zwischen Textteilen parallel erfassen kann, ist grundsätzlich gut geeignet, solche Dokumente zu analysieren.

3 Einsatz von LLMs in der finanziellen Informationsverarbeitung

Nachdem die technischen Grundlagen von Large Language Models dargestellt wurden, ist im nächsten Schritt zu klären, in welchem Ausmaß diese Modelle bereits für die Verarbeitung finanzrelevanter Informationen eingesetzt werden. Die bisherige Literatur zeigt, dass LLMs im Rechnungs- und Finanzwesen insbesondere zur Sentimentanalyse, Klassifikation, Informationsextraktion, Textgenerierung und Zusammenfassung genutzt werden. Dong et al. (2024) zeigen in ihrem Überblick, dass sich die Forschung vor allem auf Anwendungen im Asset Pricing und Investment konzentriert, bei denen aus Finanztexten, Unternehmensberichten und Marktinformationen Signale für Prognosen, Portfolioentscheidungen oder Investitionsempfehlungen abgeleitet werden. Daneben werden LLMs auch für die Klassifikation von Berichtsinhalten eingesetzt, etwa zur Identifikation von ESG-Passagen, zur Trennung von Fakten und Meinungen oder zur Ableitung von Unternehmensmerkmalen aus narrativen Berichtsteilen (Kuroki et al., 2023; W. Li et al., 2025). Darüber hinaus werden sie verwendet, um Antworten auf Earnings-Call-Fragen zu generieren, Schlüsselbegriffe zu formulieren oder Finanzberichte kompakt zusammenzufassen (Boyson et al., 2023; Dong et al., 2024; Harris, 2025). Insgesamt deutet die Literatur somit darauf hin, dass LLMs nicht nur einzelne Textfragmente verarbeiten, sondern finanzielle Informationen in einer Weise strukturieren können, die für Entscheidungsprozesse relevant sein kann.

Für die vorliegende Arbeit ist dabei besonders relevant, wie gut LLMs Finanzberichte in ihren beiden zentralen Dimensionen verarbeiten können: numerische und narrative Finanzinformationen. Numerische Informationen umfassen Kennzahlen, Bilanz- und GuV-Positionen sowie Zeitreihen, während narrative Informationen textuelle Abschnitte wie Management Discussion and Analysis, Risikoberichte oder Governance-Teile betreffen. Diese Unterscheidung ist deshalb zentral, weil die Qualität finanzieller Handlungsempfehlungen nicht allein davon abhängt, ob ein Modell Zahlen korrekt liest oder Texte sprachlich zusammenfasst. Entscheidend ist vielmehr, ob es beide Informationsarten in eine ökonomisch sinnvolle

Gesamteinschätzung überführen kann. Die folgende Literaturübersicht ordnet die bisherigen Befunde entlang dieser beiden Dimensionen und leitet daraus die Forschungslücke der Arbeit ab.

3.1 LLMs und numerische Finanzinformation

Die Literatur zur numerischen Informationsverarbeitung zeichnet zunächst ein vorsichtiges Bild. Qian et al. (2022) weisen darauf, dass große Sprachmodelle zwar bei einfacheren arithmetischen Mustern Leistungen zeigen, bei steigender Komplexität numerischer Aufgaben aber deutlich abbauen. Für den Finanzkontext konkret zeigen Zhao et al. (2024), dass leistungsfähige Modelle numerische Fragen aus Finanzdokumenten grundsätzlich beantworten können, ihre Genauigkeit jedoch sinkt, sobald Dokumente länger, tabellenreicher und inhaltlich komplexer werden. Die Autoren folgern daraus, dass LLMs zwar zu numerischem Schlussfolgern fähig sind, diese Fähigkeit in realistischen, umfangreichen Finanzdokumenten aber deutlich instabiler wird. Ergänzend zeigen neuere Arbeiten, dass Standard-LLMs weiterhin zu numerischen Halluzinationen, Problemen bei Aggregationen und Fehlern beim Retrieval aus längeren Zahlenreihen neigen, was gerade im Finanzbereich kritisch ist, da bereits kleine numerische Abweichungen ökonomisch relevante Fehlurteile nach sich ziehen können (H. Li et al., 2025).

Trotz dieser Einschränkungen sprechen auch Studien gegen die Annahme, dass LLMs generell über keine belastbare finanzielle Numerikkompetenz verfügen. Niszczota & Abbas (2023) zeigen, dass GPT-4 in einem Financial-Literacy-Test Fragen zu Zinseszins, Inflation, Diversifikation und Altersvorsorge nahezu fehlerfrei beantwortet. Darüber hinaus finden X. Li et al. (2024), dass ChatGPT bei der Prognose zentraler Unternehmenskennzahlen in mehreren Fällen näher an den realisierten Werten liegt als Analysten und dabei konservativere Einschätzungen liefert, was als Gegenpol zu einem häufig diskutierten Optimismusbias professioneller Analysten interpretiert werden kann. Y. Cheng et al. (2026) zeigen zudem, dass ChatGPT aus gegebenen Bilanz- und Erfolgsgrößen ökonomisch sinnvolle neue Faktoren generieren kann, die in Portfolioanwendungen Prognosekraft entfalten. Diese Befunde legen nahe, dass LLMs numerische Finanzinformation unter bestimmten Bedingungen nicht nur erkennen, sondern auch für quantitative Beurteilungen und Prognosen nutzen können.

Einen weiteren wichtigen Teilaspekt bildet die Extraktion strukturierter Finanzdaten aus Rechnungslegungsdokumenten. Balsiger et al. (2024) zeigen, dass GPT-4 und Bard grundsätzlich Bilanz- und GuV-Daten aus Jahresabschlüssen extrahieren können, wobei GPT-4 klar bessere Ergebnisse erzielt. Gleichzeitig treten jedoch systematische Fehler auf, insbesondere bei Größenordnungen und bei weniger standardisierten Kennzahlen wie EBIT oder Bruttogewinn. Für die vorliegende Arbeit ist dieser Befund relevant, weil er zeigt, dass die Verarbeitung von Finanzberichten bereits auf der Ebene der Identifikation und korrekten Interpretation einzelner Inhalte fehleranfällig sein kann. Wenn numerische Inhalte nicht zuverlässig erkannt oder eingeordnet werden, kann dies auch die Qualität der daraus abgeleiteten finanziellen Handlungsempfehlungen beeinträchtigen.

Insgesamt ergibt sich für die numerische Dimension ein differenziertes Bild. LLMs weisen weiterhin erkennbare Schwächen bei numerischer Präzision, komplexen Berechnungen und langen Dokumenten auf. Zugleich zeigt die Literatur aber auch, dass die Modelle grundlegende finanzielle Rechen- und Prognoseaufgaben teilweise überzeugend lösen und in bestimmten Settings sogar mit Analysten konkurrieren können.

3.2 LLMs und narrative Finanzinformation

Neben der numerischen Dimension ist die Verarbeitung narrativer Finanzinformation für die vorliegende Arbeit von zentraler Bedeutung. Bilinski (2024) und Hao (2025) betonen, dass Finanzberichte immer umfangreicher werden und insbesondere der Anteil unstrukturierter Textbestandteile stetig zunimmt. Gerade diese narrativen Teile enthalten Informationen zur Unternehmenslage, zu Risiken und zu künftigen Erwartungen, lassen aber zugleich erheblichen Spielraum für sprachliche Rahmung und strategische Selbstdarstellung. Bochkay et al. (2023) zeigen in ihrem Review, dass die Accounting-Literatur narrative Disclosures bisher vor allem über Sentiment, Lesbarkeit, Similarity und Forward-Looking Statements untersucht hat. Lange Zeit dominierten dabei Dictionary-Ansätze und andere vergleichsweise einfache NLP-Verfahren. Deren zentrale Schwäche besteht darin, dass sie Kontexte, semantische Nuancen und Bedeutungsverschiebungen oft nur unzureichend erfassen. Gerade hier wird das Potenzial moderner LLMs sichtbar, die Beziehungen zwischen Wörtern und Aussagen kontextsensitiv verarbeiten können.

Besonders relevant für Finanzberichte ist die Studie von Bilinski (2024), die zeigt, dass GPT-4 aus narrativen Teilen von Jahresberichten Sentiment- und Komplexitätsscores ableiten kann, die in signifikanter Beziehung zu Preisreaktionen und späterem ROA-Wachstum stehen. Dies spricht dafür, dass LLMs aus narrativen Berichtsinhalten nicht nur sprachliche Muster erfassen, sondern wirtschaftlich sinnvolle Informationen gewinnen können.

Auch auf Nachrichtenebene weisen LLMs ein beachtliches Potenzial auf. Kirtac & Germano (2024) zeigen, dass LLM-basierte Sentimentmaße traditionelle Dictionary-Methoden übertreffen und in Handelsstrategien ökonomisch relevante Signale liefern. Lopez-Lira & Tang (2025) kommen zu dem Ergebnis, dass GPT-4 aus bloßen Nachrichtenüberschriften die Richtung kurzfristiger Marktreaktionen und nachfolgender Renditedrift mit hoher Treffsicherheit ableiten kann, obwohl das Modell nicht speziell auf Finanzdaten feinjustiert wurde. Ergänzend zeigen D. Li et al. (2026), dass spezialisierte Architekturen für Finanzsentiment, die semantische, numerische und ereignisbezogene Informationen kombinieren, Standardmodelle wie GPT-4 sogar noch übertreffen können. Insgesamt deutet diese Forschung darauf hin, dass LLMs narrative Finanzinformationen besonders effektiv in marktrelevante Signale übersetzen können. Zugleich wird aber sichtbar, dass die Leistungsfähigkeit durch domänenspezifische Anpassungen noch weiter gesteigert werden kann.

Zusammenfassend lässt sich aus der Literatur ableiten, dass sich LLMs und konkret ChatGPT gut für die Verarbeitung narrativer Finanzinformationen eignen, weil sie Ton, Komplexität und inhaltliche Signale aus Berichten und Nachrichten so erfassen können, dass diese eng mit Kursreaktionen und zukünftiger Unternehmensentwicklung zusammenhängen.

3.3 LLMs in Finanzanalyse und Handlungsempfehlung

Über die getrennte Betrachtung numerischer und narrativer Informationen hinaus ist für die Forschungsfrage entscheidend, ob LLMs aus den verarbeiteten Informationen auch konkrete finanzielle Urteile oder Empfehlungen ableiten können. Ein Strang der Literatur untersucht dies über Aktienratings und Attraktivitätseinschätzungen. Pelster & Val (2024) zeigen in einem Live-Experiment mit S&P-500-Aktien, dass ChatGPT-4 unter Einbezug aktueller Webinformationen Attraktivitätsscores vergibt, die positiv mit künftigen Aktienrenditen korrelieren. Zudem weisen Earnings-Forecast-Ratings des Modells einen signifikanten Zusammenhang mit tatsächlichen Unternehmensgewinnen auf, besonders ausgeprägt in

Situationen, in denen Analysten untereinander stark uneinig sind. Dies deutet darauf hin, dass LLMs unter bestimmten Bedingungen Informationen verarbeiten können, die im Analysten- Konsens noch nicht vollständig eingepreist sind.

Ein weiterer Teil der Literatur betrachtet LLMs im Kontext der Portfolioerstellung und Asset-Selektion. Schneider & Yilmaz (2025) zeigen, dass GPT-Modelle Portfolios in Abhängigkeit von vorgegebenen Risikoprofilen erstellen können und dabei in einzelnen Fällen Benchmarks übertreffen. Ko & Lee (2024) kommen zu dem Ergebnis, dass ChatGPT bei der Auswahl von Assets zu stärker diversifizierten Portfolios beiträgt als Zufallsauswahlen und seine Auswahlentscheidungen mit finanztheoretisch plausiblen Überlegungen begründet. Diese Studien deuten darauf hin, dass LLMs nicht nur Informationen klassifizieren oder bewerten, sondern auch in strukturierten Entscheidungssituationen wirtschaftlich nachvollziehbare Auswahl- und Gewichtungsentscheidungen treffen können.

Darüber hinaus zeigen Arbeiten zu LLM-basierten Trading-Agenten, dass die Integration von Berichten, News und speicherbasierten Entscheidungsmechanismen in konkrete Buy-, Sell- oder Hold-Signale überführt werden kann. Yu et al. (2025) zeigen, dass ein GPT-4-basierter Trading-Agent mit geeigneter Memory-Architektur textuelle und fundamentale Informationen so verarbeiten kann, dass im untersuchten Zeitraum höhere kumulative Renditen als bei Referenzstrategien erzielt werden. Auch wenn solche Ergebnisse stark vom jeweiligen Untersuchungsdesign abhängen und nicht unmittelbar auf reale Marktanwendungen übertragbar sind, verdeutlichen sie doch, dass LLMs prinzipiell in der Lage sind, komplexe Informationsmengen in handlungsorientierte Entscheidungen zu transformieren.

Gleichzeitig zeigt sich jedoch, dass die bisherige Forschung meist eng umrissene Teilprobleme untersucht. Entweder stehen einzelne Aufgaben wie Sentimentklassifikation, Kennzahlenprognose, Portfoliobildung oder Trading im Vordergrund, oder die Modelle werden in stark kontrollierten Settings getestet. Deutlich weniger erforscht ist bislang, ob ein Modell wie ChatGPT vollständige Finanzberichte so verarbeiten kann, dass daraus Handlungsempfehlungen entstehen, die mit den Einschätzungen professioneller Finanzanalystinnen und Finanzanalysten vergleichbar sind. Gerade diese Frage ist jedoch besonders relevant, weil Analystenurteile in der Praxis typischerweise nicht auf isolierten Kennzahlen oder einzelnen Textfragmenten beruhen, sondern auf einer integrierten Bewertung von Zahlen, Narrativen und deren Implikationen für die künftige Unternehmensentwicklung.

4 Forschungsfrage und Hypothesen

Die bisherige Literatur lässt sich damit in drei Kernaussagen bündeln. Erstens zeigen die vorhandenen Studien, dass LLMs sowohl numerische als auch narrative Finanzinformationen grundsätzlich verarbeiten können, wobei ihre Schwächen insbesondere in der numerischen Präzision, in komplexen Rechenoperationen und in der Verarbeitung langer, tabellenreicher Dokumente liegen. Zweitens belegen zahlreiche Arbeiten, dass die aus diesen Informationen gewonnenen Signale ökonomisch relevant sein können, etwa für Marktreaktionen, Kennzahlenprognosen, Portfolioselektion oder Handelsstrategien. Drittens fehlt bislang weitgehend eine Untersuchung, die direkt prüft, ob ein Modell wie ChatGPT vollständige Finanzberichte so analysieren kann, dass daraus Handlungsempfehlungen entstehen, die mit Analystenkonsens mithalten und sich an der tatsächlichen Kursentwicklung messen lassen.

Genau an dieser Forschungslücke setzt die vorliegende Arbeit an. Während bisherige Studien meist isolierte Aufgaben oder stark spezialisierte Anwendungsszenarien betrachten, untersucht diese Arbeit die Fähigkeit von ChatGPT, vollständige Finanzberichte in einer Weise zu verarbeiten, die zu mit Experten vergleichbaren finanziellen Handlungsempfehlungen führt.

Vor diesem Hintergrund stellt sich die zentrale empirische Frage dieser Arbeit, inwiefern ChatGPT als Instrument zur finanziellen Informationsverarbeitung und als Ratgeber bei Anlageentscheidungen mit dem Konsens professioneller Finanzanalysten mithalten kann. Konkret sollen Handlungsempfehlungen von Analysten und ChatGPT basierend auf Quartalsberichten miteinander verglichen und anhand des tatsächlichen späteren Kursverlaufs bewertet werden.

Um diese Fragestellung zu beantworten, wird im Folgenden ein quantitatives Untersuchungsdesign beschrieben, das Handlungsempfehlungen von ChatGPT auf Basis von 10-Q-Berichten erhebt und diese Empfehlungen mit dem Konsens von Analysten und der tatsächlichen Kursentwicklung in einem Event-Study-Rahmen verknüpft.

Einhergehend mit der Forschungsfrage werden 2 Hypothesen aufgestellt:

- H1: Basierend auf Quartalsberichten unterscheiden sich Handlungsempfehlungen der Analysten von jenen von ChatGPT

- H2 ChatGPT gibt weniger zuverlässige Handlungsempfehlungen (kurzfristig sowie mittel- bis langfristig) blickend auf den späteren Aktienverlauf als Analysten

Hypothese 2 basiert auf der Annahme, dass Analysten gegenüber ChatGPT eine überlegene Prognosefähigkeit besitzen. Zwar wird im vorliegenden Untersuchungsdesign beiden Seiten dieselbe formale Informationsbasis, die jeweiligen Quartalsberichte, zugrunde gelegt, jedoch verfügen Analysten in der Praxis über ein wesentlich breiteres Informationsspektrum. Dieses umfasst etwa branchenspezifisches Expertenwissen, historische Unternehmenseinschätzungen, Managementgespräche sowie laufend aktualisierte Markt- und Wettbewerbsinformationen. ChatGPT hingegen wird im Rahmen dieser Studie explizit angewiesen, seine Empfehlung ausschließlich auf Basis des bereitgestellten Quartalsberichts abzugeben. Diese strukturelle Informationsasymmetrie zugunsten der Analysten bildet die zentrale Begründung für die in Hypothese 2 formulierte Erwartung einer höheren Treffsicherheit der menschlichen Experten.

5 Daten und Methodik

Im Folgenden wird zunächst das empirische Setting der Arbeit im Überblick dargestellt. Anschließend werden die zentralen Bausteine der Methode in getrennten Unterabschnitten vertieft sowie konkret, wie diese Studie durchgeführt wird.

Die Arbeit verfolgt ein quantitatives Forschungsdesign, das die Eignung von ChatGPT als Instrument zur finanziellen Informationsverarbeitung und als Ratgeber bei Anlageentscheidungen mit der Prognosegüte professioneller Finanzanalysten vergleicht. Kern der Untersuchung sind 50 Unternehmen aus dem S&P-500-Index, für die nach der Veröffentlichung eines Quartalsberichts (Form 10-Q) sowohl der Analystenkonsens als auch die Handlungsempfehlung von ChatGPT erhoben und mit der tatsächlichen späteren Kursentwicklung der jeweiligen Aktie abgeglichen werden. Auf Basis dieser Daten wird zum einen geprüft, ob sich die Niveaus der Empfehlungen (Ratingskala 1–5) zwischen Analysten und ChatGPT signifikant unterscheiden, und zum anderen, welche Seite, menschliche Analysten oder das LLM, im Hinblick auf kurz-, mittel- und langfristige Kursbewegungen häufiger „richtiger“ liegt. Dazu werden die Empfehlungen beider Seiten in Richtungskategorien übersetzt, mit der Kursentwicklung in verschiedenen Zeitfenstern nach den Empfehlungen verglichen und über eine Score-Logik mit kumulierten Renditen bewertet.

5.1 Stichprobe

Die Stichprobe setzt sich aus 50 nach Marktkapitalisierung gereihten Unternehmen des S&P-500 zum Stichtag 30.09.2025 zusammen. Mit dem Fokus auf große, liquide Indexwerte wird eine hinreichende Analystenabdeckung sowie verlässliche Kurs- und Fundamentaldaten gewährleistet.

Für jedes dieser Unternehmen wird der erste Quartalsbericht vom Typ Form 10-Q herangezogen, der nach dem Trainings-Cutoff von GPT-5.4 (August 2025) veröffentlicht wurde.

Im Rahmen von Hypothese 2 wird ein Beobachtungszeitraum von 120 Handelstagen nach Veröffentlichung des relevanten 10-Q-Berichts gesetzt. Da die Geschäftsjahre der Unternehmen unterschiedlich verlaufen, liegt der erste 10-Q-Report nach dem Trainings-Cutoff von GPT-5.4 bei einigen Firmen so spät im Jahr, dass zwischen diesem Veröffentlichungsdatum und dem Stichtag der Renditeauswertung (08.05.2026) kein Beobachtungsfenster von 120 Handelstagen gegeben ist. Solche Fälle werden aus der Stichprobe entfernt und durch Unternehmen aus der nachfolgenden Rangliste ersetzt, für die die vollständige Beobachtungsperiode vorliegt. Auf diese Weise umfasst die endgültige Stichprobe weiterhin 50 S&P-500-Unternehmen, stellt aber sicher, dass für alle Beobachtungen der gewünschte Renditehorizont von 120 Handelstagen abgedeckt ist. Konkret werden dabei die Unternehmen Nvidia, Apple, Visa, Walmart, Costco, Home Depot, Cisco und Walt Disney aus der Stichprobe entfernt und mit Thermo Fischer Scientific, American Express, Booking Holdings, Lam Research, Blackrock, GE Vernova, Texas Instruments und Boeing ersetzt. Für das Unternehmen Alphabet ist zu berücksichtigen, dass im S&P 50 zwei Aktienklassen gelistet sind, die sich auf denselben Quartalsbericht beziehen. In dieser Studie wird ausschließlich die Klasse-A-Aktie berücksichtigt; die Klasse-C-Aktie bleibt außen vor, um Doppelzählungen desselben Unternehmens zu vermeiden und die Stichprobe auf eine wirtschaftliche Einheit pro Firma zu beschränken.

Als Informationsbasis dienen, wie bereits erwähnt, standardisierte Form-10-Q-Berichte, die US-börsennotierte Unternehmen für die ersten drei Quartale eines Geschäftsjahres bei der Securities and Exchange Commission (SEC) einreichen müssen. Diese Zwischenberichte enthalten eine Bilanz, Gewinn- und Verlustrechnung, Kapitalflussrechnung sowie umfangreiche Erläuterungen (Notes) und einen Management Discussion & Analysis-Teil, in dem Kennzahlen kommentiert, Strategien erläutert und Risiken beschrieben werden (SEC, o. J.).

Die 10-Q-Berichte sind weniger umfassend als der Jahresbericht (Form 10-K), folgen aber einer klaren regulatorischen Struktur, was die Vergleichbarkeit zwischen Unternehmen erhöht. Für diese Arbeit sind 10-Q-Berichte besonders geeignet, weil sie neben den quantitativen Finanzkennzahlen auch umfangreiche qualitative Informationen enthalten. Während davon auszugehen ist, dass Analysten in ihrer Bewertung zusätzlich auf ein breites Spektrum externer Informationsquellen zurückgreifen, ist ChatGPT in diesem Studiendesign theoretisch auf den Inhalt des jeweiligen Quartalsberichts beschränkt. Gerade deshalb ist es vorteilhaft, dass der 10-Q nicht nur Zahlen, sondern auch narrative Abschnitte wie Management Diskussionen, Risikoberichte und Erläuterungen zu strategischen Entwicklungen umfasst. Diese qualitativen Passagen erlauben es dem Modell, über reine Zahleninterpretation hinaus Kontext, Unternehmensumfeld und zukünftige Chancen und Risiken zu berücksichtigen und sich dadurch inhaltlich einem Informationsstand anzunähern, der dem von Analysten zumindest teilweise vergleichbar ist.

Die Quartalsreports werden von der öffentlich zugänglichen Datenbank der Securities and Exchange Commission SEC Edgar Datenbank bezogen. Die Reports werden jeweils in Form vom HTML-Dateityp hochgeladen, weshalb die in ChatGPT eingespielten Quartalsberichte auch in diesem Format in den Chats hochgeladen werden.

Eine Studie von Kovorova-Simacek et al. (2026) zeigt, dass große Sprachmodelle strukturierte, maschinenlesbare Berichte in HTML-Format deutlich häufiger als Primärquelle heranziehen, aus ihnen zuverlässiger Fakten extrahieren und dadurch seltener auf externe Drittquellen ausweichen müssen als zum Beispiel bei eingespielten PDF-Berichten. Damit ist die Verwendung von Quartalsberichten im HTML-Format für diese Studie vorteilhaft und kann die Genauigkeit der GPT-generierten Auswertungen begünstigen.

Abschließend, waren für die endgültige Stichprobe noch einige Anpassungen erforderlich. Neben jenen Unternehmen, deren Quartalsberichte zu spät veröffentlicht wurden und kein Beobachtungsfenster von 120 US-Handelstagen ermöglichten, wurden JPMorgan und Goldman Sachs entfernt, da ihre 10-Q-HTML-Dateien laut GPT-Fehlermeldung zu umfangreich waren, sowie Bank of America, Wells Fargo, Morgan Stanley und Citigroup, bei denen wiederholt ein unbekannter Verarbeitungsfehler auftrat, der vermutlich ebenfalls auf die Dateigröße zurückzuführen ist. In allen Fällen erfolgte der Ersatz systematisch durch nach Indexgewicht nachrückende Unternehmen, deren 10-Q-Report vom Modell verarbeitet werden konnte und deren Veröffentlichungsdatum ein 120-tägiges Beobachtungsfenster zuließ. Potenzielle Nachrücker wie TJX oder Accenture konnten beispielsweise nicht berücksichtigt

werden, da ihre erste 10-Q-Veröffentlichung nach dem Trainings-Cutoff des Modells zu spät erfolgte.

5.2 Datenerhebung ChatGPT

Bei der Datenerhebung auf ChatGPT-Seite wird zunächst der Trainings-Cutoff des Modells berücksichtigt. Die verwendete ChatGPT-Version 5.4 verfügt laut öffentlich zugänglicher Dokumentation über einen Wissensstand, der bis 31. August 2025 reicht (OpenAI, o. J.-b). Somit sind spätere Entwicklungen, insbesondere Kursverläufe nach dem dritten Quartal 2025, nicht im Modellgedächtnis enthalten. Indem für jedes Unternehmen der erste 10-Q nach dem 31. August 2025 als Informationsbasis gewählt wird, ist sichergestellt, dass ChatGPT die nachfolgenden Kursbewegungen nicht einfach aus den Trainingsdaten „wiedererkennt“, sondern seine Einschätzung tatsächlich auf der Interpretation des konkreten Berichts aufbaut. Diese Vorgehensweise wird in den Studien von X. Li et al. (2024) und Schneider & Yilmaz, (2025) gehandhabt, die ebenfalls Vorhersagen von ChatGPT für die Zukunftsperiode nach dem Trainings-Cutoff einholen.

Dabei wird in dieser Arbeit explizit GPT-5.4 Thinking verwendet. Im Abschnitt 2.2.1 wurde die Thinking Variante bereits erklärt. Für die vorliegende Untersuchung scheint die Wahl der Thinking-Variante am sinnvollsten, da die Analyse eines vollständigen 10-Q-Berichts nicht nur eine schnelle Antwort, sondern die strukturierte Verarbeitung umfangreicher quantitativer und qualitativer Informationen erfordert. Außerdem nutzt diese Variante die volle Leistungsfähigkeit der GPT-5.4-Architektur und stellt damit die leistungsstärkste verfügbare Repräsentation des Modells dar, sodass der Vergleich mit den Analystenempfehlungen auf der qualitativ höchsten GPT-Benchmark erfolgt. Zu erwähnen sei, dass zum Zeitpunkt der Durchführung dieser Studie die Thinking-Variante nicht für die kostenlose Nutzung von ChatGPT verfügbar ist.

Wichtig zu erwähnen sei noch, dass die Studie im normalen Chat-User-Interface von ChatGPT durchgeführt wird und keine API-Anbindung verwendet wird. Mittels APIs wäre eine automatisierte und skalierbarere Arbeitsweise möglich, jedoch gehen damit auch direkte Konfigurationsmöglichkeiten einher, Parameter wie die Temperatur des Modells manuell zu setzen, womit das Modell in eine gewisse Richtung gelenkt wäre. Die Nutzung des Standard-User-Interface hat den Vorteil, dass das Modell unter den von OpenAI vorgesehenen

Standardbedingungen betrieben wird und keine manuelle Beeinflussung der Sampling-Parameter stattfindet. Dadurch spiegeln die generierten Empfehlungen das Verhalten von ChatGPT wider, wie es auch einem regulären Nutzer zugänglich ist, was die praxisnähere Einschätzung der Leistungsfähigkeit des Modells ermöglicht.

5.2.1 Prompting

Die Empfehlungen von ChatGPT werden durch kontrolliertes Prompting erhoben. Der verwendete Prompt lautet:

„Bitte analysiere den folgenden Quartalsbericht (Form 10-Q) eines Unternehmens. Deine Aufgabe ist es, ausschließlich basierend auf den im Quartalsbericht enthaltenen Informationen eine Handlungsempfehlung für die Aktie des Unternehmens abzugeben. Der Quartalsreport liegt im HTML-Format vor.

Die Handlungsempfehlung soll auf folgender Skala erfolgen:

1 = Strong Buy
2 = Buy
3 = Hold
4 = Sell
5 = Strong Sell

Die Antwort soll ausschließlich aus einer einzelnen Zahl der erwähnten Skala bestehen (1–5).“

Mit dieser Promptgestaltung werden mehrere Ziele verfolgt. Erstens wird generell ein klarer und präziser Prompt verwendet, der dem Modell eine eindeutig definierte Aufgabe sowie eine klar formulierte Antwortskala vorgibt, wie es Chen et al. (2025) empfehlen. Zweitens wird die Informationsbasis bewusst auf den bereitgestellten Bericht beschränkt, indem explizit formuliert wird, dass die Empfehlung „ausschließlich“ auf den Informationen des 10-Q beruhen soll. Buergler et al. (2025), welche ebenfalls Output von ChatGPT basierend auf hochgeladenem 10-K filing einholen, verwenden die Formulierung „The provided 10-K filing is the whole content for addressing my question“ um Aufmerksamkeit des Modells auf das relevante Dokument zu setzen. Ein analoges Prinzip kommt hier zum Einsatz, um zusätzliche, außerhalb des Reports liegende Wissensbestände möglichst auszuschließen. Drittens

wird die Input-Semantik präzisiert, indem darauf hingewiesen wird, dass der Bericht im HTML-Format vorliegt. Die Arbeit von White et al. (2023) legt nahe, dass die Beschreibung des Dateityps und Kontextes für die interne Interpretation durch das Modell hilfreich ist, ohne das Ergebnis direkt zu steuern. Zugleich wird beim Prompting darauf geachtet, keine unnötigen Zusatzinstruktionen oder irrelevanten Inhalte mitzuliefern, um keinen künstlichen Informations-Overload zu erzeugen, der die Aufmerksamkeit des Modells von den zentralen Informationen im Report ablenken könnte.

Bewusst verzichtet die Studie auf die verbreiteten Prompt-Engineering-Techniken few-shot-prompting und Chain-of-Thought (CoT) Prompting. Beide Techniken sind zwar grundsätzlich geeignet, die Leistung von LLMs zu steigern, greifen aber maßgeblich in den Entscheidungsprozess des Modells ein. Chain-of-Thought-Prompts können dem Modell indirekt vorgeben, über welche gedanklichen Zwischenschritte es seine Antwort entwickeln soll, während Few-shot-Prompts über Demonstrationsbeispiele implizit ein bestimmtes Antwortmuster und eine „Erwartungshaltung“ vorzeichnen. Um die Neutralität des Experiments zu wahren und die Vergleichbarkeit mit Analystenempfehlungen, die ebenfalls nicht entlang explizit vorgegebener „Beispielratings“ entstehen, sicherzustellen, werden Few-shot und CoT daher gezielt nicht eingesetzt. Der in dieser Studie eingesetzte Prompt entspricht also einem zero-shot-prompting.

Bezüglich der Technik des Role Promptings, bei der dem Modell eine bestimmte Rolle zugewiesen wird (z.B. „You are a financial advisor“), kann dieses zwar ein nützliches Werkzeug sein, um Antworten stärker zu konditionieren, hat sich in Tests zur finanziellen Grundkompetenz jedoch nicht zwingend als vorteilhaft erwiesen. In der Studie zur Finanzkompetenz von GPT von Niszczota & Abbas (2023) war die Performance von GPT bei einer Rollenzuweisung als Finanzberater schlechter als bei der Variante ohne dem Role-Prompting. X. Li et al. (2024), die ChatGPT Prognosen mit jenen von Experten vergleichen, verwenden kein explizites Role-Prompting, sondern arbeiten mit inhaltlich klaren, aber rollenneutralen Anweisungen. Vor diesem Hintergrund wird in der Hauptanalyse auf Role Prompting verzichtet, um die Modellantwort auch nicht in eine künstliche Rolle zu drängen.

Es wird jedoch ein ergänzender Durchlauf mit einer Rollen-Zuweisung getestet und analysiert, ob diese die Empfehlungen systematisch verändert. Immerhin haben Schneider & Yilmaz (2025) bei der Portfolioerstellung durch ChatGPT mit „Pretend to be a financial expert“ dem Modell eine Rolle zugewiesen. Auch bei Magner et al. (2025) und Yu et al. (2025) gab es eine Rollenzuteilung für die KI. Im Durchlauf dieser Arbeit wird GPT die Rolle eines

Aktienanalysten zugewiesen. Der bereits vorgestellte Prompt wird dabei lediglich dahingehen verändert, dass als erster Satz „Übernimm die Rolle eines professionellen Aktienanalysten.“ hinzugefügt wird.

Wichtig ist zu erwähnen, dass sowohl beim Durchlauf ohne der Rollenzuweisung als auch mit, für jedes Unternehmen der Prompt (samt Quartalsreport) jeweils in 3 unabhängigen Chats wiederholt und der Mittelwert der ausgegebenen Zahlen als GPT-Rating herangezogen wird. Mit dieser Vorgehensweise wird die im Kapitel 2.1 thematisierte stochastische Natur des Modells berücksichtigt, die unterschiedliche Ergebnisse in unterschiedlichen Chats verursachen kann. Dieser sogenannten Re-Sampling Ansatz wird auch in der Studie von Li et al. (2024) verfolgt, die in ihrem ähnlichen Studien-Design viele separate und unabhängige Chat-Durchläufe ausführen und die Ergebnisse dann zusammenaggregieren.

5.3 Datenerhebung Analysten

Auf der Analystenseite bildet eine professionelle Finanzdatenbank die Grundlage der Datenerhebung. LSEG Workspace (vormals Eikon with Datastream) ist ein umfassendes Finanzinformations- und Analysetool, das aktuelle und historische Unternehmens- und Marktdaten bereitstellt, darunter auch Konsens-Analystenempfehlungen und Gewinnschätzungen. Für jedes in die Stichprobe einbezogene Unternehmen und jede 10-Q-Veröffentlichung wird aus LSEG Workspace die tägliche Konsensbewertung der Analysten auf der Skala von 1 (Strong Buy) bis 5 (Strong Sell) abgerufen. Diese Konsensratings aggregieren die Bewertungen verschiedener Analystenhäuser und liefern damit einen marktweiten Expertenkonsens.

Die Konsensempfehlung wird am dritten Handelstag nach der 10-Q-Veröffentlichung (t= +3) als Referenzrating verwendet, um eine realistische Verarbeitungszeit der Analysten zu berücksichtigen. Beccalli et al. (2015) untersuchen, wie Analysten Informationen konkret vom Mikroprozessor-Unternehmen Intel verarbeiten und stellen fest, dass für klassische, periodische Finanzinformationen wie Quartals- und Jahresberichte Analysten ihre EPS-Schätzungen typischerweise innerhalb von 0 bis 3 Kalendertagen nach Veröffentlichung des Berichts anpassen. Y. Li et al. (2020) zeigen, dass die Reaktionszeit von Analysten davon abhängt, wie vollständig die begleitenden Finanzinformationen bereits in der Earnings-Pressemeldung vor der offiziellen 10-Q-Einreichung offengelegt werden. Ist der Anteil der 10-Q-

Informationen hoch, die erst mit dem 10-Q veröffentlicht werden und nicht schon in der Pressemeldung verkündet werden, verschiebt sich ein größerer Teil der Analystenreaktionen von der unmittelbaren Earnings-Pressemeldung hin zu Zeitpunkten nach der 10-Q-Einreichung, während bei niedriger Verzögerung ein Großteil der Forecasts bereits innerhalb von zwei Handelstagen nach der Earnings-Meldung erfolgt. Das impliziert, dass Analysten ihre Schätzungen grundsätzlich sehr zeitnah anpassen und selbst bei verzögerter Informationsoffenlegung des Unternehmens spätestens kurz nach der 10-Q-Veröffentlichung aktualisieren, sodass ein enges Reaktionsfenster um die Earnings-Ankündigung und die nachfolgende 10-Q-Einreichung den wesentlichen Teil der Schätzungsanpassungen erfasst.

Im Endeffekt besteht kein eindeutiges, einheitliches „Reaktionsfenster". Die Wahl von drei Handelstagen stellt basierend auf der erwähnten Literatur einen plausiblen Kompromiss dar, der einerseits eine gewisse Anpassungszeit zulässt und andererseits den kausalen Bezug zur 10-Q-Disclosure nicht überdehnt. Um diese Annahme abzusichern, wird ein Robustheitscheck durchgeführt, in dem die Konsensratings zu $t+1$, $t+2$, $t+3$ und $t+5$ nach der 10-Q-Veröffentlichung miteinander verglichen und analysiert werden.

Für die Bewertung der Prognosegüte der Empfehlungen werden unterschiedliche Prognosefenster definiert. Jegadeesh et al. (2004) analysieren den Informationsgehalt von Konsensempfehlungen und Empfehlungsänderungen über Quartals-Horizonte hinweg und messen die Performance typischerweise über Holding-Period-Returns ab dem ersten Handelstag des Folgequartals, also über Zeiträume von mehreren Wochen bis einigen Monaten. Desai et al. (2000) untersuchen die Performance von Analysten-Empfehlungen mit Buy-and-Hold-Renditen zwischen 10 und 500 Tagen und zeigen, dass sich sowohl sehr kurzfristige Reaktionen als auch längerfristige Entwicklungen sinnvoll über diskrete Halteperioden abbilden lassen. In Anlehnung an diese Studien werden für die vorliegende Arbeit mehrere Beobachtungsfenster gesetzt, die jeweils am ersten Handelstag nach Vorliegen der konsolidierten Konsensempfehlung beginnen, also ab $t+4$, da ein rationaler Investor die zum Zeitpunkt $t+3$ vorliegende Empfehlung praktisch oft frühestens am folgenden Handelstag umsetzen kann. Konkret werden ein sehr kurzfristiges Fenster von 10 Handelstagen (≈ 2 Wochen), ein kurz- bis mittelfristiges Fenster von 60 Handelstagen (≈ 3 Monate) sowie ein mittleres bis längeres Fenster von 120 Handelstagen (≈ 6 Monate) nach $t+4$ betrachtet. Diese Wahl lehnt sich an die bei Desai et al. (2000) verwendeten 10-Tage- und mehrmonatigen Buy-and-Hold-Perioden an und spiegelt zugleich den von Jegadeesh et al. (2004) dokumentierten quartalsweisen Bewertungshorizont von Empfehlungen wider.

Aufgrund der zeitlichen Beschränkung durch den Knowledge-Cutoff von GPT-5.4 (August 2025) und den Beobachtungszeitraum der Kursdaten, der zum Zeitpunkt dieser Arbeit nur bis zu Mai 2026 reicht, ist ein 500-Tage- bzw. volles 12-Monats-Fenster nicht möglich, stattdessen markieren 120 Handelstage das maximal ausgewertete „längerfristige“ Intervall. Außerdem zielen zwar Analystenempfehlungen wahrscheinlich auf einen mittleren bis längeren Anlagehorizont, werden in der Praxis aber laufend an neue Informationen angepasst und sind nicht zum Beispiel für einen starren Zwölfmonatszeitraum fixiert. Deshalb scheint es methodisch auch sinnvoll, das längste Beobachtungsfenster nicht zu weit in die Zukunft zu strecken. Sehr lange Horizonte würden die Wirkung des ursprünglichen, auf den betrachteten Quartalsbericht bezogenen Signals mit späteren Prognoseänderungen und neuen Nachrichten vermischen. Die Beschränkung auf maximal rund 120 Handelstage ermöglicht dagegen, die Prognosegüte näher an der ursprünglichen Informationsbasis, nämlich dem 10-Q-Report und der dazugehörigen Empfehlung zu bewerten und Überlagerungen durch spätere Ereignisse zu begrenzen.

In einer neueren Studie, in der Kadam & Sethi (2024) am indischen Aktienmarkt Analystenprognosen mit dem tatsächlichem späterem Aktienkursverlauf vergleichen, verwenden die Autoren ein 12-Monats-Fenster und messen, dass rund 63 Prozent der Zielkurse innerhalb eines Jahres und 38% der Kursziele genau zum Ende des 12-Monats-Horizonts erreicht oder überschritten werden. Damit scheint die Wahl des Beobachtungsfensters von bis zu 120 Handelstagen in dieser Arbeit sinnvoll, da sich die Beobachtung innerhalb der Zeitspanne der Prognoserealisierungen befindet und die Kurszielrealisationen bei Kadam & Sethi (2024) gegen Ende des 12-Monats-Horizonts sogar abnehmen.

Zusammenfassend lässt sich festhalten, dass die Methodik zunächst ein überblicksartig beschriebenes Setting etabliert: 50 S&P-500-Unternehmen, 10-Q-Berichte als zentraler Informationsschock, ChatGPT- und Analystenempfehlungen als konkurrierende Ratgeber, und Kursverläufe als objektiver Bewertungsmaßstab.

5.4 Methode

Zur Prüfung von Hypothese 1 werden die von Analysten und GPT-5.4 Thinking erzeugten Konsens-Handlungsempfehlungen zunächst auf der identischen numerischen Skala von 1 bis 5 abgebildet. Für jedes Unternehmen liegt damit ein Wertepaar aus Analystenrating und GPT-Rating vor, jeweils zum gleichen Referenzzeitpunkt $t = +3$ Handelstage nach Veröffentlichung des relevanten Quartalsberichts. Auf Basis dieser Paare werden im ersten Schritt deskriptive Auswertungen vorgenommen, etwa Mittelwerte, Mediane, Standardabweichungen und grafische Darstellungen, um einen Überblick über das Niveau und die Verteilung der Bewertungen beider Seiten zu erhalten. Im zweiten Schritt wird die Hypothese, dass sich die durchschnittlichen Ratings unterscheiden, mit einem gepaarten T-Test untersucht. Der Test erfolgt auf einem Signifikanzniveau von 5%, sodass die Nullhypothese „kein systematischer Unterschied zwischen den durchschnittlichen Ratings“ verworfen wird, wenn der beobachtete Mittelwertunterschied statistisch nicht mehr plausibel allein durch Zufallsschwankungen erklärbar ist.

Zur Prüfung von Hypothese 2 wird die Prognosegüte der Empfehlungen von Analysten und GPT anhand eines Richtigkeitsscores bewertet. Dazu werden die auf einer fünfstufigen Ratingskala erhobenen Empfehlungen beider Seiten zunächst in Richtungskategorien übersetzt und mit der realisierten Rendite der jeweiligen Aktie in mehreren vordefinierten Halteperioden verglichen. Die fünfstufige Ratingskala (1-5) wird zu den Kategorien „positiv“, „neutral“ und „negativ“ zusammengefasst. Konsens-Empfehlungen kleiner als 2,5 werden als positive Einschätzung (Kaufempfehlung), Konsens-Empfehlungen im Bereich 2,5 bis exklusive 3,5 als neutrale Einstufung und Konsens-Empfehlungen größer gleich 3,5 werden als negative Einschätzung (Verkaufs- bzw. Untergewichtungsempfehlung) interpretiert. Die tatsächliche Kursentwicklung in den definierten Beobachtungsfenstern wird anhand der realisierten Rendite in ein entsprechendes Vorzeichen übersetzt, sodass für jede Aktie und jedes Zeitfenster geprüft werden kann, ob die Richtung der Empfehlung zur späteren Kursbewegung passt.

Die Zeitfenster sind wie im Datenerhebungs-Abschnitt beschrieben: 10, 30, 60 und 120 US-Handelstage nach dem Referenzzeitpunkt. Der Referenzzeitpunkt ist dabei der Konsenszeitpunkt $t + 3$ Handelstage nach der 10-Q-Veröffentlichung ($t = 0$). Es wird angenommen, dass ein typischer Investor zum Aktien-Schlusskurs an $t + 3$ auf Basis der vorliegenden Konsensempfehlung kauft bzw. verkauft. Die Rendite beginnt dann ab dem folgenden

Handelstag $t + 4$ zu laufen und wird über die jeweiligen Halteperioden von 10, 30, 60 und 120 Handelstagen gemessen. Somit entsprechen die Beobachtungsfenster genauer dargestellt $t = 4$ bis 13, bis 33, bis 63 sowie bis $t = 123$. Dieser Ansatz wird auch in Desai et al. (2000) verfolgt, hier wird der Aktien-Schlusskurs auch zum Tag der Veröffentlichung der Analysten-Empfehlungen für die Renditeberechnungen herangezogen und die Beobachtungsfenster starten ab dem nächsten Tag nach Veröffentlichung der Analysten-Empfehlungen.

Auf Basis der Richtungen von den Empfehlungen und Kursbewegungen wird ein diskreter Richtigkeitsscore vergeben, der die Übereinstimmung zwischen Empfehlung und Kursverlauf abbildet. Liegt die Kursentwicklung in derselben Richtung wie die Empfehlung (z.B. positive Empfehlung und tatsächlich positive Rendite bzw. negative Empfehlung und tatsächlich negative Rendite), wird ein Treffer mit einem +1 Score bewertet. Verfehlt die Empfehlung die Richtung der Kursbewegung (z.B. Kaufempfehlung bei anschließender negativer Rendite), erhält die Beobachtung keinen Scorerpunkt (=0). Die so konstruierten Scores werden für jedes Unternehmen und jedes Renditefenster getrennt für Analysten und GPT berechnet. Auf beiden Seiten werden Trefferquoten ermittelt und die Unterschiede zwischen diesen mittels gepaartem T-Test wieder auf einem Signifikanzniveau von 5% untersucht.

Für die Hauptanalyse zu Hypothese 2 wird die Empfehlungskategorie „Hold“ bei der Vergabe der Richtigkeitsscores nicht berücksichtigt. In einem zusätzlichen Robustheitscheck wird „Hold“ jedoch einem spezifischen neutralen Renditeband zugeordnet und damit in die Score-Vergabe einbezogen. Diese zweite Auswertung dient dazu zu prüfen, ob die Ergebnisse der Hauptanalyse gegenüber einer alternativen Behandlung der „Hold“-Kategorie stabil bleiben. In diesem Robustheitscheck bleibt die Definition der Richtigkeitsscores für Kauf- und Verkaufsempfehlungen unverändert, das heißt Buy-Ratings gelten als korrekt, wenn die Rendite im betrachteten Fenster positiv ist, und Sell-Ratings, wenn die Rendite negativ ist. Abweichend von der Hauptanalyse wird jedoch die Kategorie „Hold“ explizit berücksichtigt, indem sie als korrekt gewertet wird, wenn die Rendite in ein neutrales Renditeband zwischen dem 40. und 60. Perzentil der jeweiligen insgesamten Fenster-Renditen aller Unternehmen der Stichprobe fällt, und als inkorrekt, wenn die Rendite außerhalb dieses Bereichs liegt. Die Wahl der Spanne zwischen dem 40. und 60. Perzentil hat den Hintergrund, dass auf der verwendeten Ordinalskala von 1 bis 5 die Kategorie ‚Hold‘ (= 3) ebenfalls nur 20% des Ratingspektrums einnimmt.

Um die Wahl der Reaktionszeit von drei Handelstagen zu bekräftigen, wird wieder ein Robustheitscheck durchgeführt. Dabei werden die Konsensempfehlungen der Analysten zu 1, 2, 4 und 5 Handelstagen nach Veröffentlichung des jeweiligen 10-Q-Berichts ($t = 1$, $t = 2$, $t = 4$, $t = 5$) erhoben und mit den Konsensratings zu $t = 3$ verglichen. Zu diesem Zweck werden gepaarte t-Tests zwischen den Ratings zu $t = 3$ und den jeweiligen Alternativzeitpunkten durchgeführt, um zu prüfen, ob sich die Mittelwerte der Konsensempfehlungen signifikant unterscheiden.

Für den Buy- und Sell-Bereich werden sowohl Strong Buy und Buy als auch Strong Sell und Sell als positive beziehungsweise negative Kaufempfehlungen einheitlich dahingehend bewertet, ob der Kurs anschließend tatsächlich gestiegen oder gefallen ist. In der Scorer-Vergabe gibt es also zum Beispiel keinen Unterschied zwischen Strong Buy und Buy bei positiver Aktienentwicklung. Um die feiner abgestufte Informationsstruktur der Skala dennoch zu nutzen, werden die Buy/Sell Bereiche zusätzlich separat auf GPT- und Analystenseite ausgewertet. Für Strong Buy wird ein Intervall bis exklusive 1,5 sowie für Buy ein Intervall von 1,5 bis unter 2,5 definiert, sinngemäß für Strong Sell (im Intervall über 4,5) und Sell (Intervall zwischen 3,5 bis unter 4,5). Für alle 4 Untergruppen werden eigene Trefferquoten berechnet.

Für alle vier Ratingstufen innerhalb des Buy- bzw. Sell-Bereichs werden separate, bedingte Trefferquoten berechnet, die angeben, wie häufig eine Empfehlung einer bestimmten Stufe zu einer prognosekonformen Kursentwicklung geführt hat (z.B. Anteil der Fälle, in denen ein Strong-Sell-Rating von einer tatsächlich negativen Kursrendite gefolgt wurde). Zusätzlich werden für jede dieser Ratingstufen die durchschnittlichen kumulierten Renditen in den jeweiligen Prognosehorizonten ausgewiesen, sodass neben der reinen Häufigkeit richtiger Einschätzungen auch die typische Stärke der nachfolgenden Kursbewegungen sichtbar wird. Durch den Vergleich dieser Trefferquoten und Durchschnittsrenditen zwischen Strong Buy und Buy sowie zwischen Strong Sell und Sell wird deskriptiv untersucht, ob besonders starke positive bzw. negative Konsensempfehlungen mit überdurchschnittlich hohen Kursgewinnen bzw. -verlusten einhergehen oder ob sich nur geringe Unterschiede zu moderaten Kauf- oder Verkaufsempfehlungen zeigen. Die Untergruppen dienen darüber hinaus dazu, für jede Ratingstufe GPT-Empfehlungen und Analystenkonsense direkt nebeneinanderzustellen, um etwa zu prüfen, ob GPT-Strong-Buys im Durchschnitt ähnlich häufig und ähnlich stark „treffen“ wie Strong-Buy-Konsense der Analysten oder ob sich systematische Unterschiede in den beobachteten Trefferraten und Renditen ergeben.

Ergänzend wird die Differenz der Empfehlungen, definiert als Diff_Rating = $\text{Rating}_{GPT} - \text{Rating}_{Analysten}$, in einem ersten Regressionsmodell als abhängige Variable verwendet, um zu analysieren, ob Niveauunterschiede zwischen GPT und Analysten mit firmenspezifischen Merkmalen wie Sektorzugehörigkeit, Marktkapitalisierung um den Berichtszeitpunkt und historischer Volatilität zusammenhängen. In einem zweiten Regressionsmodell dient die Differenz der Richtigkeitsscores von GPT und Analysten als abhängige Variable, sodass positive Werte anzeigen, dass GPT in einem gegebenen Fenster präziser lag als der Analystenkonsens, und negative Werte auf eine relative Unterlegenheit der KI hinweisen. Durch die Aufnahme derselben Kontrollvariablen wie im ersten Modell wird untersucht, ob die Prognosegüte von GPT gegenüber Analysten in bestimmten Segmenten, etwa in bestimmten Branchen, Größenklassen oder Risikoprofilen, systematisch höher oder niedriger ausfällt.

Konkret werden in den beiden Regressionsanalysen folgende Kontrollvariablen berücksichtigt, die jeweils zum Zeitpunkt des 10-Q-Events bzw. im vordefinierten Vorlaufzeitraum ab dem jeweiligen Handelstag vor der Veröffentlichung der Quartalsreports erhoben werden.

- Unternehmensgröße: Als Unternehmensgröße wird der Marktwert des Eigenkapitals (Market Cap in USD) ein US-Handelstag vor der 10-Q-Veröffentlichung entnommen. Zusätzlich wird das Index-Gewicht nach Marktkapitalisierung im S&P 500 des jeweiligen Unternehmens berücksichtigt.
- Bewertungskennzahlen: Kurs-Gewinn pro Aktie-Verhältnis (Price / Earnings per Share) und Kurs-Buchwert pro Aktie-Verhältnis (Price / Book Value per Share) entnommen aus LSEG Workspace zum Stichtag einen Handelstag vor der Veröffentlichung des Quartalsberichts, um finanzielle Fundamentaldaten der Unternehmen zu erfassen. Im weiteren Verlauf der Arbeit werden diese Variablen vereinfacht als Kurs-Gewinn-Verhältnis (KGV) und Kurs-Buchwert-Verhältnis (KBV) bezeichnet. Anzumerken ist, dass die Unternehmen McDonald‘s, Booking, Boeing und Phillip Morris laut LSEG Workspace KBV-Werte von „NULL“ einen Handelstag vor den jeweiligen 10-Q Disclosures vorweisen. Dies ist darauf zurückzuführen, dass die Buchwerte dieser Unternehmen zu diesen Zeitpunkten negativ waren und deswegen kein Verhältnis ausgerechnet werden kann. In die Regressionsanalyen werden die KBV-Werte dieser Unternehmen daher auch als 0 aufgenommen. Bei Boeing wird zudem auch ein 0-Wert zum KGV zugewiesen, da die Firma auch einen negativen Gewinn (=Verlust) zum gemessenen Zeitpunkt einfuhr.

- Historische Unternehmens-Volatilität: Standardabweichung der täglichen Aktienrenditen in einem 60-Tage-Fenster vor der 10-Q-Veröffentlichung. Konkret werden für die Berechnung zunächst die täglichen, logarithmischen Preisänderungen von Tag zu Tag ermittelt. Aus diesen relativen historischen Preisänderungen wird im nächsten Schritt die Standardabweichung berechnet. Diese tägliche Standardabweichung wird anschließend auf das gesamte Jahr hochgerechnet (annualisiert) und das Endergebnis in Prozent ausgedrückt. Diese Berechnung folgt jener von LSEG Workspace für historische Volatilitäten. Für diese Arbeit repräsentiert diese Variable vereinfacht das aktuelle Risikoprofil des jeweiligen Unternehmens.
- Unternehmens-Momentum: kumulierte Aktienrendite über einen mehrmonatigen Zeitraum von 60 US-Handelstagen vor der 10-Q Veröffentlichung, um aktuelle Trendbewegungen der jeweiligen Aktie zu kontrollieren. Die Rendite werden jeweils mit der Formel (Pt-1 / Pt-60) – 1 ermittelt.
- Industrie-Dummies: Indikatorvariablen auf Basis der Industriezugehörigkeit in LSEG Workspace, um Unterschiede zwischen Industrien zu untersuchen.

  Die ursprüngliche Industrievariable enthielt insgesamt 30 verschiedene, sehr spezifische Branchenklassifikationen. Um die Anzahl der Dummy-Variablen zu reduzieren und die Analyse übersichtlicher zu gestalten, wurden diese einzelnen Branchen zu 10 übergeordneten Industriekategorien zusammengefasst. Dafür wurden inhaltlich ähnliche Branchen gebündelt. Beispielsweise wurden Software, Halbleiter, IT-Dienstleistungen und elektronische Komponenten der Kategorie Technology_industry zugeordnet, während Pharmaunternehmen, Medizintechnik und Managed Healthcare in der Kategorie Healthcare_industry zusammengefasst wurden. Energiebezogene Branchen wie Öl- und Gasunternehmen sowie Versorgungsunternehmen wurden zu Energy_Utilities_industry gebündelt. Konsum- und Einzelhandelsbranchen wurden in Consumer_Retail_industry eingeordnet, während Finanzdienstleistungen wie Investment Management und Investment Banking in Financial_Services_industry zusammengeführt wurden. Durch diese Umkodierung wurde die ursprüngliche, sehr detaillierte Branchenvariable vereinfacht, ohne die grundlegende wirtschaftliche Zugehörigkeit der Unternehmen zu verlieren.

  Die jeweiligen Dummy-Variablen nehmen den Wert 1 an, wenn ein Unternehmen der jeweiligen Industrie zugeordnet ist, und ansonsten den Wert 0. Als

Referenzkategorie dient der Sektor „Technology_industry“, da dieser mit 12 Beobachtungen den relativ größten Anteil in der Stichprobe stellt und damit eine empirisch gut besetzte und zugleich inhaltlich zentrale Vergleichsbasis bildet. Die geschätzten Koeffizienten der übrigen Sektoren sind somit jeweils als Abweichung vom Technologiesektor zu interpretieren.

- Markt-Momentum: kumulierte Rendite des S&P-500-Index in einem Zeitfenster von 60 US-Handelstagen vor dem 10-Q-Bericht, um die allgemeine Marktstimmung zu erfassen. Die Berechnung erfolgt gleicherweise wie beim Unternehmens-Momentum, nur werden statt Preisänderungen der jeweiligen Aktien die jeweiligen S&P-Rendite herangezogen.
- Marktvolatilität: Niveau der Marktvolatilität vor dem Berichtsstichtag, errechnet durch die Standardabweichung der täglichen Index-Renditen des S&P 500 60 Handelstage vor der Quartalsberichtserstattung, um Phasen erhöhter Unsicherheit und Marktturbulenzen als mögliche Einflussfaktoren zu kontrollieren. Die Berechnung erfolgt auf gleicher Weise wie bei der historischen Unternehmens-Volatilität mit dem Unterschied, dass Index-Preisänderungen statt Aktien-Preisänderungen berücksichtigt werden.
- Dokumentengröße der Quartalsberichte (in Kilobyte), die als Maß für den technischen Informationsumfang dient, der ChatGPT zur Verfügung steht.
- Durchschnittliche Nachdenkzeit von GPT in Sekunden. Hierfür wird je Unternehmen der Mittelwert der Nachdenkzeiten über alle drei GPT-Durchläufe gebildet.

Alle diese Kontrollvariablen haben als Referenzzeitpunkt $t = -1$, also ein Handelstag vor der Veröffentlichung der jeweiligen 10-Q's. Die 60-Tage-Fenster in der Vergangenheit strecken sich also von $t = -1$ bis $t = -60$ aus. Marktkapitalisierung und Bewertungskennzahlen werden zu $t = -1$ erhoben. Auf diese Weise spiegeln diese Kontrollvariablen das Unternehmens- und Marktumfeld wider, das zum Zeitpunkt der Informationsveröffentlichung (10-Q) vorliegt, ohne bereits durch Reaktionen auf den Bericht und die darauffolgenden Empfehlungen beeinflusst zu sein. Ergänzend werden mit der Dokumentengröße und der durchschnittlichen Nachdenkzeit potenzielle Einflussfaktoren auf der GPT-Seite berücksichtigt.

6 Ergebnisse

6.1 Unterschied zwischen GPT- und Analystenkonsens

Für Hypothese 1 wurden zunächst die Verteilungen der Konsensempfehlungen von Analysten und GPT beschrieben.

Die GPT-Konsens-Ratings reichen in der Stichprobe von 1,33 bis 4, während die niedrigste Konsens-Rating-Vergabe bei den Analysten 1,69 und die höchste 2.85 beträgt. Der Median des Analystenkonsenses liegt bei 2,09, während der Median des GPT-Konsenses bei 2,00 liegt. 25% der GPT-Konsens-Empfehlungen befinden sich bei einem Rating von bis zu 2 und 75% bis zum Rating 2,67. Bei den Analysten beträgt das 1. Quartil 1,94 und das 3. Quartil 2,28.

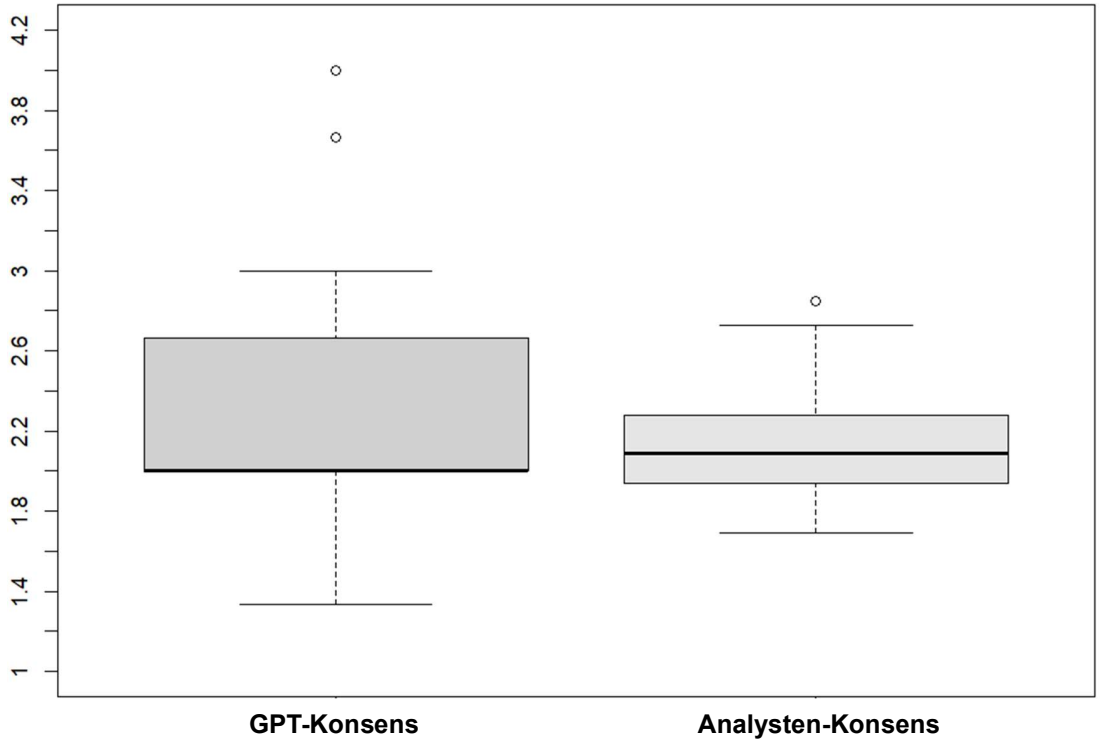
Die Mittelwerte betragen 2,14 für den Analystenkonsens und 2,29 für den GPT-Konsens, mit Standardabweichungen von 0,28 (Analysten) bzw. 0,55 (GPT). Insgesamt deuten diese Kennzahlen darauf hin, dass die GPT-Empfehlungen im Durchschnitt etwas höher ausfallen als die Empfehlungen der Analysten und somit GPT relativ konservativer beziehungsweise die Analysten im Durchschnitt relativ optimistischer bewerteten.

In Tabelle 1 ist die soeben erläuterte deskriptive Statistik sowie in Abbildung 1 die jeweiligen Boxplots beider Gruppen dargestellt.

Tab. 1: Konsensverteilungen Übersicht

	1. Quartil	Median	3. Quartil	Mittelwert	Standardabweichung
GPT-Konsens	2	2	2,67	2,29	0,55
Analysten-Konsens	1,94	2.09	2,28	2,14	0,28

Abb. 1: Boxplots - Konsensverteilungen



Zur formalen Prüfung der Mittelwertunterschiede zwischen den gepaarten Beobachtungen wurde ein gepaarter t-Test auf die Differenz der Konsensempfehlungen (GPT minus Analystenkonsens) durchgeführt. Der Mittelwert der Differenz beträgt 0,1497, der Test ergab einen t-Wert von 1,9556 bei 49 Freiheitsgraden und einen p-Wert von 0,0562. Das zugehörige 95%-Konfidenzintervall für die mittlere Differenz liegt zwischen −0,0041 und 0,3040.

Der p-Wert überschreitet das konventionelle Signifikanzniveau von 5%, liegt jedoch unterhalb eines weniger strengen Signifikanzniveaus von 10%, sodass der Mittelwertunterschied nach dem t-Test auf dem 5%-Niveau nicht, auf dem 10%-Niveau jedoch als statistisch signifikant einzustufen ist.

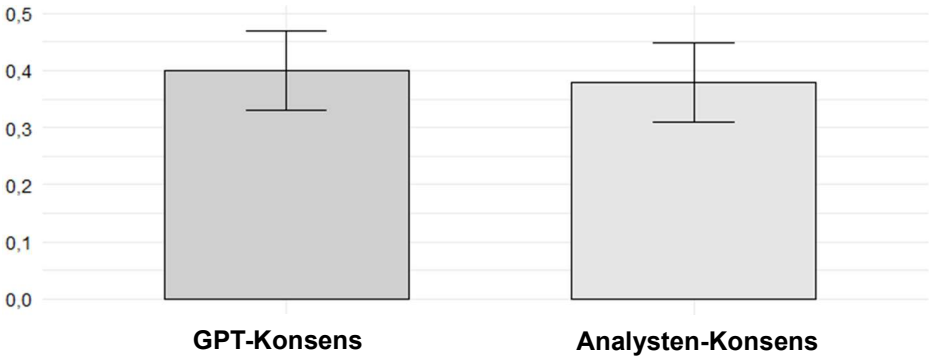
Die Handlungsempfehlungen von ChatGPT unterscheiden sich damit von jenen der Analysten (auf Basis der Quartalsberichte), der Unterschied ist jedoch knapp nicht auf dem 5%-Signifikanzniveau statistisch abgesichert. Somit wird Hypothese 1, dass sich die beiden Empfehlungskonsense voneinander unterscheiden, als nicht bestätigt eingestuft.

6.2 Trefferquotenvergleich

Für Hypothese 2 wurden die Richtigkeitsscores von GPT und Analystenkonsens in vier Renditefenstern nach Veröffentlichung des 10-Q-Berichts betrachtet (10, 30, 60 und 120 Handelstage).

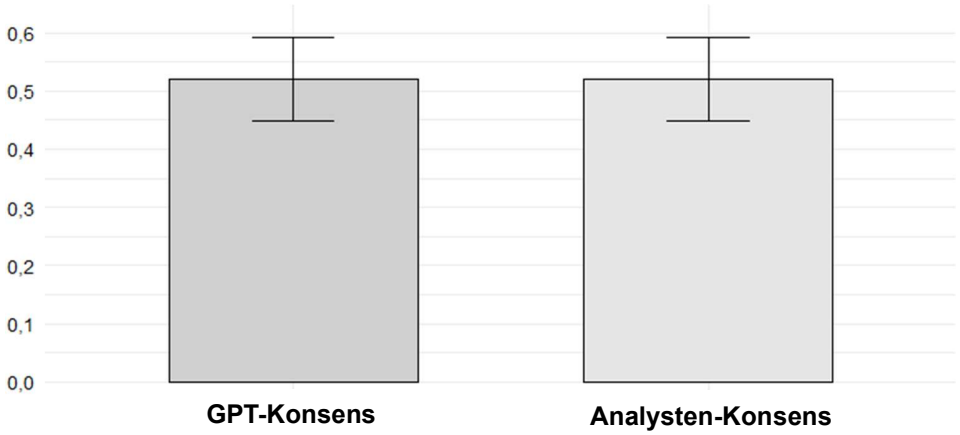
Im 10-Tage-Fenster (t13) beträgt die Trefferquote der Analysten 0,38, während GPT in 40% der Fälle korrekt lag. Die Standardabweichungen von 0,49 (Analysten) bzw. 0,4949 (GPT) zeigen in beiden Gruppen eine ähnliche Streuung der Treffer-Scores, und die Standardfehler von 0,0693 bzw. 0,0700 deuten auf eine vergleichbare Präzision der geschätzten Trefferquoten hin. Abbildung 2 stellt die Trefferquoten visuell dar.

Abb. 2: Trefferquoten nach 10 Tagen



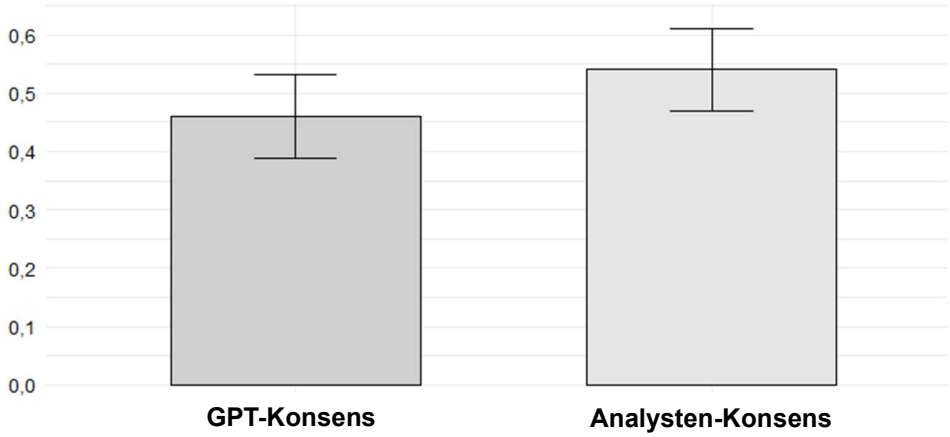
Im 30-Tage-Fenster (t33) erreichen sowohl Analysten als auch GPT eine Trefferquote von 0,52, bei identischen Standardabweichungen von 0,5047 und Standardfehlern von jeweils 0,0714. Damit sind sowohl Mittelwerte als auch Streuungsmaße in diesem Fenster praktisch gleich. Abbildung 3 bestätigt dieses Muster visuell.

Abb. 3: Trefferquoten nach 30 Tagen



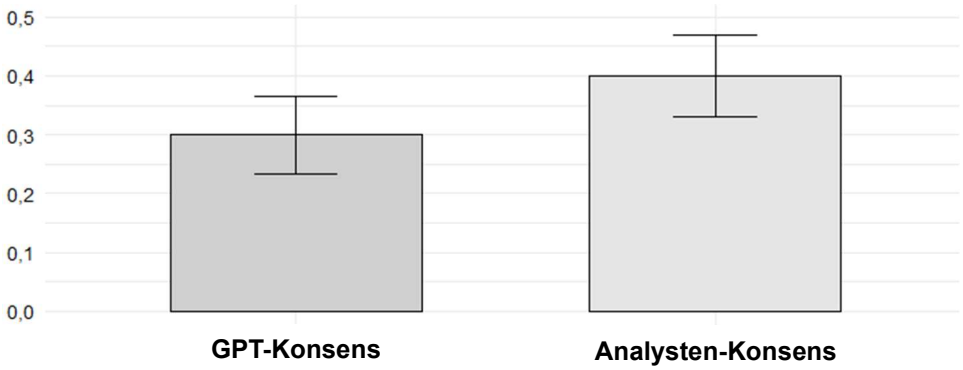
Im 60-Tage-Fenster (t63), wie aus Abbildung 4 zu entnehmen, liegt die Trefferquote der Analysten bei 0,54, während GPT in 46% der Fälle korrekt ist. Trotz dieses Niveauunterschieds sind Standardabweichungen (jeweils 0,5035) und Standardfehler (0,0712) erneut nahezu identisch.

Abb. 4: Trefferquoten nach 60 Tagen



Im 120-Tage-Fenster (t123) beträgt die Trefferquote der Analysten 0,40 gegenüber 0,30 bei GPT. Die Standardabweichung ist bei den Analysten mit 0,4949 etwas höher als bei GPT (0,4629), die Standardfehler von 0,0700 (Analysten) und 0,0655 (GPT) weisen auf eine ähnlich genaue Schätzung der jeweiligen Trefferquoten hin. In Abbildung 5 sind die Trefferquoten beider Seiten visuell dargestellt.

Abb. 5: Trefferquoten nach 120 Tagen



Zur formalen Prüfung der Unterschiede in den Trefferquoten wurden für jedes Renditefenster gepaarte t-Tests durchgeführt, die die mittlere Differenz zwischen GPT-Treffern und Analystentreffern pro Unternehmen auswerten. Im 10-Tage-Fenster (t13) liegt die mittlere Differenz (GPT minus Analysten) bei 0,02. Der p-Wert von 0,7664 zeigt, dass der beobachtete Unterschied sehr klein ist und statistisch klar nicht von null zu unterscheiden ist. Das 95%-Konfidenzintervall von −0,1145 bis 0,1545 umfasst null deutlich, was ebenfalls darauf hinweist, dass sich die durchschnittlichen Trefferquoten in diesem Fenster nicht signifikant unterscheiden. Im 30-Tage-Fenster (t33) beträgt die mittlere Differenz 0,00, der t-Wert von 0,0, der p-Wert von 1,0 sowie das symmetrische Konfidenzintervall von −0,1519 bis 0,1519 unterstreichen, dass es hier keinen messbaren Unterschied zwischen GPT und Analysten gibt. Für das 60-Tage-Fenster (t63) ergibt sich eine mittlere Differenz von −0,08 zugunsten der Analysten, p-Wert von 0,2522 zeigt jedoch, dass auch dieser Unterschied statistisch nicht signifikant ist, was durch das Konfidenzintervall von −0,2188 bis 0,0588 (das null einschließt) bestätigt wird. Im 120-Tage-Fenster (t123) liegt die mittlere Differenz bei −0,10, der t-Wert von −1,6977 und der p-Wert von 0,0959 deuten auf einen etwas größeren, aber weiterhin nur grenzwertigen Unterschied hin. Das 95%-Konfidenzintervall von −0,2184 bis 0,0184 schließt null knapp ein, sodass auch für das längste Beobachtungsfenster kein statistisch gesicherter Unterschied auf dem üblichen 5%-Signifikanzniveau festgestellt wird.

Zusammenfassend zeigt sich, dass die Trefferquoten von GPT und Analystenkonsens über die betrachteten Renditefenster hinweg relativ nahe beieinander liegen, wobei GPT kurzfristig (10 Handelstage) leicht höhere und die Analysten in den längeren Zeitfenstern (60 und 120 Handelstage) tendenziell höhere Trefferquoten aufweisen. Keiner dieser Unterschiede ist jedoch statistisch signifikant auf dem üblichen 5%-Signifikanzniveau, sodass sich

für keines der Renditefenster ein klar überlegener Ansatz in Bezug auf die Trefferwahrscheinlichkeit nachweisen lässt.

Auf Basis dieser Ergebnisse wird Hypothese 2 nicht bestätigt. ChatGPT gibt damit kurz- wie auch mittel- bis langfristig auf Basis von Quartalsberichten nicht weniger zuverlässige Handlungsempfehlungen ab als der Analystenkonsens.

6.3 Robustheitscheck; Inklusion von „Hold“

Für den Robustheitscheck mit Einbezug der Kategorie „Hold“ wurden erneut die Trefferquoten von GPT und Analystenkonsens in den vier Renditefenstern (10, 30, 60 und 120 Handelstage) betrachtet.

Bei 10 Handelstagen (t13) liegt die Trefferquote der Analysten bei 0,40, während GPT eine leicht höhere Trefferquote von 0,42 erreicht. Die Standardabweichungen betragen 0,4949 (Analysten) bzw. 0,4986 (GPT), die Standardfehler 0,0700 bzw. 0,0705 und weisen damit auf eine sehr ähnliche Streuung und Schätzpräzision in beiden Gruppen hin.

Nach 30 Handelstagen (t33) beträgt die Trefferquote der Analysten 0,52, während GPT mit 0,58 eine höhere Trefferquote erzielt. Die Standardabweichungen liegen hier bei 0,5047 (Analysten) und 0,4986 (GPT), die Standardfehler bei 0,0714 bzw. 0,0705.

Im 60-Tage-Fenster (t63) weisen die Analysten eine Trefferquote von 0,56 auf, während GPT in 48% der Fälle korrekt ist. Die Standardabweichungen von 0,5014 (Analysten) und 0,5047 (GPT) sowie die Standardfehler von 0,0709 bzw. 0,0714 sind erneut sehr ähnlich.

Im längsten Fenster von 120 Handelstagen (t123) liegt die Trefferquote der Analysten bei 0,42 gegenüber 0,34 bei GPT. Die Standardabweichungen betragen 0,4986 (Analysten) und 0,4785 (GPT), die Standardfehler 0,0705 bzw. 0,0677 und deuten auch hier auf vergleichbare Streuungen hin.

Insgesamt zeigen die deskriptiven Kennzahlen, dass sich die Trefferquoten von GPT und Analysten durch die Berücksichtigung von „Hold“ zwar leicht verschieben (GPT liegt im 10- und 30-Tage-Fenster etwas über, im 60- und 120-Tage-Fenster etwas unter den Analysten) die Größenordnungen und Streuungen bleiben jedoch über alle Fenster hinweg ähnlich.

Zur formalen Prüfung der Unterschiede in den modifizierten Treffer-Scores wurden wie in der Hauptanalyse gepaarte t-Tests zwischen GPT und Analystenkonsens für jedes Renditefenster durchgeführt.

Im 10-Tage-Fenster (t13) beträgt die mittlere Differenz (GPT minus Analysten) 0,02. Der t-Wert von 0,2988 (df = 49) und der p-Wert von 0,7664 zeigen, dass der beobachtete Unterschied sehr gering ist und statistisch klar nicht von null zu unterscheiden ist. Das 95%-Konfidenzintervall reicht von −0,1145 bis 0,1545. Nach 30 Handelstagen (t33) liegt die mittlere Differenz bei 0,06 zugunsten von GPT, der t-Wert von 0,9029 und der p-Wert von 0,371 deuten ebenfalls auf keinen statistisch signifikanten Unterschied hin, was durch das 95%-Konfidenzintervall von −0,0736 bis 0,1935 bestätigt wird. Im 60-Tage-Fenster (t63) ergibt sich eine mittlere Differenz von −0,08 zugunsten der Analysten, der t-Wert von −1,2727 (df = 49) und der p-Wert von 0,2091 zeigen, dass auch dieser Unterschied statistisch nicht signifikant ist. Das Konfidenzintervall von −0,2063 bis 0,0464 schließt null ein. Für das 120-Tage-Fenster (t123) beträgt die mittlere Differenz −0,08, der t-Wert liegt bei −1,1586 (df = 49) und der p-Wert bei 0,2522. Das 95%-Konfidenzintervall von −0,2188 bis 0,0588 umfasst ebenfalls null und weist damit auf keinen statistisch gesicherten Unterschied hin.

Insgesamt zeigen die t-Tests, dass auch bei Einbezug der „Hold“-Kategorie in der Scorer-Vergabe in keinem der betrachteten Renditefenster ein statistisch signifikanter Unterschied in den Trefferquoten zwischen GPT und Analystenkonsens festgestellt werden kann, sodass die zentralen Ergebnisse der Hauptanalyse durch diesen Robustheitscheck bestätigt werden.

6.4 Robustheitscheck; Reaktionszeit der Analysten

Die Robustheitsanalyse der Reaktionszeiten zeigt, dass sich der Analystenkonsens über die fünf betrachteten Handelstage nach Veröffentlichung des 10-Q-Berichts praktisch nicht verändert. Die Mittelwerte der Konsensempfehlungen liegen sehr eng beieinander zwischen 2,147 (t1), 2,145 (t2), 2,1436 (t3), 2,144 (t4) und 2,1428 (t5), auch die Standardabweichungen bewegen sich in einem sehr schmalen Bereich von rund 0,285 (t1) bis 0,283 (t5), was auf eine über die Zeitpunkte hinweg nahezu konstante Verteilung der Ratings hinweist. Die gepaarten t-Tests, mit denen die Konsensempfehlungen zu $t = 3$ jeweils mit $t = 1$, $t = 2$, $t = 4$ und $t = 5$ verglichen wurden, bestätigen dieses Bild: Weder der Vergleich t3 vs. t1

(mittlere Differenz = −0,0034, p = 0,139) noch t3 vs. t2 (mittlere Differenz = −0,0014, p = 0,312) weist einen statistisch signifikanten Unterschied auf. Gleiches gilt für t3 vs. t4 (mittlere Differenz = −0,0004 bei einem p-Wert von 0,5325) und t3 vs. t5 (mittlere Differenz = 0,0008 bei einem p-Wert von 0,4383). Insgesamt liefern die deskriptiven Kennzahlen und die gepaarten t-Tests somit keine Evidenz für systematische Verschiebungen des Analystenkonsenses innerhalb der ersten fünf Handelstage nach dem Quartalsbericht, sodass die Wahl von $t = 3$ als Referenzzeitpunkt für die Hauptanalysen als stabil erscheint.

6.5 Analyse feinerer Ratingstufen

Im Folgenden werden die bedingten Trefferquoten und die durchschnittlichen kumulierten Renditen der vier Ratingstufen (Strong Buy, Buy, Sell und Strong Sell) für GPT und den Analystenkonsens getrennt nach Prognosehorizont sowie aggregiert über alle Zeitfenster hinweg deskriptiv ausgewiesen. Ziel ist es, zu untersuchen, ob sich die Prognosequalität zwischen den Ratingstufen unterscheidet und ob GPT- und Analystenempfehlungen dabei ähnliche Muster aufweisen.

Da in der vorliegenden Stichprobe sowohl auf GPT- als auch auf Analystenseite die Ratingstufen Strong Buy und Strong Sell nur sehr selten oder gar nicht vergeben wurden (auf Analystenebene entsprach keine einzige Beobachtung Strong Buy oder Strong Sell, auf GPT-Seite wurde Strong Buy lediglich einmal und Strong Sell kein einziges Mal vergeben) sind die entsprechenden Trefferquoten und Renditen nicht aussagekräftig. Die nachfolgenden Ausführungen konzentrieren sich dementsprechend auf die deutlich besser besetzten Kategorien Buy und Sell.

Innerhalb der Buy-Kategorie zeigt Tabelle 2, dass GPT mit 35 und die Analysten mit 43 Buy-Empfehlungen über ähnlich viele Beobachtungen verfügen. Im kürzesten Prognosehorizont von 13 Handelstagen erzielt GPT eine Trefferquote von 48,6% bei einer durchschnittlichen Rendite von −1,1%. Die Analysten liegen mit 44,2% Trefferquote und einer mittleren Rendite von −0,9% auf vergleichbarem Niveau. Im 33-Tage-Fenster steigen die Trefferquoten beider Gruppen deutlich an: GPT trifft zu 62,9% bei einer durchschnittlichen Rendite von nahezu +0,03%, die Analysten zu 60,5% bei +0,4% mittlerer Rendite. Im 63-Tage-Fenster setzt sich dieser positive Trend fort, GPT erzielt 60,0% Trefferquote bei +0,9% mittlerer Rendite, die Analysten 62,8% bei +3,4% , bevor im längsten Horizont von 120 Handelstagen beide

Gruppen wieder zurückfallen. GPT trifft hier die Buy-Empfehlungen nur noch zu 40,0% bei einer mittleren Rendite von −0,6% und die Analysten zu 46,5% bei +2,7%.

Auffällig ist, dass bei den Buy-Empfehlungen im 60-Tage-Fenster die Analystentrefferquoten und Durchschnittsrenditen besonders günstig ausfallen, was darauf hindeutet, dass Buy-Empfehlungen der Analysten in diesem mittleren Zeithorizont die höchste praktische Relevanz entfalten. Auf GPT-Seite ist ein ähnlicher, wenngleich schwächerer Effekt erkennbar. Im 120-Tage-Fenster hingegen sind die durchschnittlichen Renditen für GPT-Buy-Empfehlungen trotz noch nennenswerter Trefferquoten leicht negativ, was darauf hindeutet, dass selbst korrekt prognostizierte Aufwärtsbewegungen im langen Horizont im Mittel nur marginal ausgeprägt sind.

Die Sell-Kategorie ist mit lediglich drei Beobachtungen auf GPT-Seite und null auf Analystenseite äußerst dünn besetzt, sodass die ausgewiesenen Kennzahlen mit Vorsicht zu interpretieren sind. Die durchschnittlichen Renditen der GPT-Sell-Empfehlungen sind in allen Zeitfenstern negativ (10 Tage: −3,6%; 30 Tage: −4,6%; 60 Tage: −4,1%; 120 Tage: −2,0%), was zumindest in die erwartete Richtung einer Kursabnahme weist, jedoch aufgrund der geringen Fallzahl keine belastbaren Schlüsse erlaubt.

Tab. 2: feinere Darstellung der Rating-Stufen pro Zeitfenster

(Nach Zeitfenstern)		GPT-Konsens			Analysten-Konsens		
t	Rating	Anzahl	Treffer ⌀	Rendite ⌀	Anzahl	Treffer ⌀	Rendite ⌀
13	Strong-Buy	1	0	-0,1405	0	n.a	n.a
13	Buy	35	0,4857	-0,01080	43	0,4419	-0,0088
13	Sell	3	1	-0,0363	0	n.a	n.a
13	Strong-Sell	0	n.a	n.a	0	n.a	n.a
33	Strong-Buy	1	1	0,1184	0	n.a	n.a
33	Buy	35	0,6286	0,0003	43	0,6047	0,0043
33	Sell	3	1	-0,046	0	n.a	n.a
33	Strong-Sell	0	n.a	n.a	0	n.a	n.a
63	Strong-Buy	1	0	-0,3755	0	n.a	n.a
63	Buy	35	0,6000	0,0093	43	0,6279	0,0342
63	Sell	3	0,6667	-0,0412	0	n.a	n.a
63	Strong-Sell	0	n.a	n.a	0	n.a	n.a

(Nach Zeitfenstern)		GPT-Konsens			Analysten-Konsens		
t	Rating	Anzahl	Treffer ⌀	Rendite ⌀	Anzahl	Treffer ⌀	Rendite ⌀
123	Strong-Buy	1	0	-0,2659	0	n.a	n.a
123	Buy	35	0,4000	-0,0060	43	0,4651	0,02722
123	Sell	3	0,3333	-0,0197	0	n.a	n.a
123	Strong-Sell	0	n.a	n.a	0	n.a	n.a

Anmerkung: Der Spaltenname „t“ steht für die Zeitreihe; t = 13 sind 10 Handelstage, t = 63 Handelstage, etc.

In der Gesamtbetrachtung über alle Prognosehorizonte hinweg (Tabelle 3) bestätigt sich das Bild für die Buy-Kategorie: GPT erzielt eine aggregierte Trefferquote von 52,9% bei einer mittleren Rendite von −0,2%, die Analysten von 53,5% bei +1,4%. Beide Gruppen liegen auf ähnlichem Trefferquotenniveau, wobei die Analysten im Durchschnitt leicht positive Renditen erwirtschaften, während GPT-Buy-Empfehlungen über alle Horizonte hinweg im Mittel marginal im negativen Bereich liegen. Auf GPT-Seite weist die Sell-Kategorie aggregiert eine Trefferquote von 75,0% und eine mittlere Rendite von −3,6% aus, ist jedoch aufgrund der minimalen Fallzahl nicht sinnvoll mit den Analysten vergleichbar.

Tab. 3: feinere Darstellung der Rating-Stufen (gesamt)

(Gesamt)	GPT-Konsens			Analysten-Konsens		
Rating	Anzahl	Treffer ⌀	Rendite ⌀	Anzahl	Treffer ⌀	Rendite ⌀
Strong-Buy	1	0,25	-0,1659	0	n.a	n.a
Buy	35	0,5286	-0,0018	43	0,5349	0,0142
Sell	3	0,7500	-0,0358	0	n.a	n.a
Strong-Sell	0	n.a	n.a	0	n.a	n.a

Insgesamt lässt sich somit festhalten, dass Analysten im Buy-Bereich in dieser Stichprobe leicht überlegen sind, während GPT im Sell-Bereich überhaupt erst verwertbare, wenn auch nur schwach prognosekräftige Signale liefert.

6.6 Regressionsanalyse

6.6.1 1. Regressionsmodell

Zur Analyse der Determinanten der Ratingdifferenz zwischen GPT und dem Analystenkonsens wird ein multiples lineares Regressionsmodell geschätzt, in dem als erklärende Variablen zentrale Unternehmens- und Marktmerkmale sowie Dummy-Variablen für neun der zehn zusammengefassten Industriesektoren (mit dem Technologiesektor als Referenzkategorie) aufgenommen werden.

Das Differenzmodell erklärt mit einem multiplen R^2 von 0,608 rund 61% der Varianz der Ratingdifferenz, während das adjustierte R^2 bei 0,36 liegt. Der Gesamt-F-Test ist mit F(19, 30) = 2,45 und einem p-Wert von 0,014 auf dem 5-Prozent-Niveau statistisch signifikant. Der Residualstandardfehler beträgt 0,433 bei 30 Freiheitsgraden. Das Regressionsmodell ist in Tabelle 4 dargestellt.

Auf Ebene der einzelnen Koeffizienten zeigen sich nur wenige statistisch signifikante Zusammenhänge. Die Marktvolatilität des S&P 500 weist mit einem Koeffizienten von −35,78 einen auf dem 5-Prozent-Niveau signifikant negativen Effekt auf (p = 0,010). Ein Anstieg der Marktvolatilität um einen Prozentpunkt geht mit einer Verringerung der Ratingdifferenz um etwa 0,358 Einheiten einher.

Das Kurs-Gewinn-Verhältnis zeigt einen schwach negativen Effekt mit einem Koeffizienten von −0,002 (p = 0,054). Steigt das KGV um eine Einheit, sinkt die Ratingdifferenz um etwa 0,002 Einheiten. Das Unternehmensmomentum weist auf dem 10-Prozent-Niveau einen positiven Koeffizienten von 1,021 auf (p = 0,093). Ein Anstieg des Unternehmensmomenturms um einen Prozentpunkt geht mit einer Erhöhung der Ratingdifferenz um etwa 0,010 Einheiten einher. Die übrigen Unternehmens- und Marktmerkmale leisten keinen statistisch nachweisbaren eigenständigen Erklärungsbeitrag.

Bei den Branchendummies zeigen sich für drei Sektoren statistisch signifikante Abweichungen gegenüber dem Referenzsektor Technologie, alle auf dem 5-Prozent-Niveau. Im Sektor Automotive/Industrial Goods liegt die Ratingdifferenz im Durchschnitt um 0,528 Einheiten höher (p = 0,038), im Sektor Healthcare um 0,478 Einheiten höher (p = 0,032) und im Sektor Energy/Utilities um 0,622 Einheiten höher (p = 0,040). In allen drei Fällen weichen GPT-Rating und Analystenkonsens also systematisch stärker voneinander ab als im Referenzsektor

Technologie. Die verbleibenden Branchendummies zeigen keine statistisch signifikanten Abweichungen.

Ergänzend zum Differenzmodell wurden zwei weitere Modelle geschätzt, bei denen das GPT-Rating bzw. der Analystenkonsens jeweils als eigenständige abhängige Variable dienen.

Das Modell mit dem GPT-Konsens als abhängiger Variable erreicht ein multiples R^2 von 0,627 und ein adjustiertes R^2 von 0,370, der Gesamt-F-Test ist mit $F(20, 29) = 2{,}437$ und einem p-Wert von 0,014 auf dem 5-Prozent-Niveau signifikant. Der Residualstandardfehler beträgt 0,433. Als signifikante Einflussfaktoren zeigen sich der Analystenkonsens selbst mit einem positiven Koeffizienten von 0,715 ($p = 0{,}018$). Steigt der Analystenkonsens um eine Einheit auf der Ratingskala, erhöht sich das GPT-Rating im Durchschnitt ebenfalls um 0,715 Einheiten in dieselbe Richtung, was auf eine substanzielle Gleichläufigkeit beider Bewertungen hindeutet. Darüber hinaus weist die Marktvolatilität des S&P 500 einen stark negativen Koeffizienten von −37,95 auf ($p = 0{,}008$). Ein Anstieg der Marktvolatilität um einen Prozentpunkt geht mit einer Verringerung des GPT-Ratings um etwa 0,380 Einheiten einher, das heißt GPT vergibt bei höherer Marktvolatilität tendenziell optimistischere Ratings. Das Unternehmensmomentum zeigt einen positiven, auf dem 10-Prozent-Niveau schwach signifikanten Koeffizienten von 1,084 ($p = 0{,}077$). Bei den Branchendummies erreichen Healthcare (0,493, $p = 0{,}028$) und Energy/Utilities (0,596, $p = 0{,}050$) das 5-Prozent-Niveau, Automotive/Industrial Goods liegt mit 0,498 ($p = 0{,}051$) knapp darüber, in allen drei Sektoren vergibt GPT relativ zum Technologiesektor höhere, also pessimistischere Ratings.

Das Modell mit dem Analystenkonsens als abhängiger Variable fällt demgegenüber deutlich schwächer aus: Das multiple R^2 beträgt 0,520, das adjustierte R^2 nur 0,189, und der Gesamt-F-Test ist mit $F(20, 29) = 1{,}571$ und einem p-Wert von 0,131 nicht statistisch signifikant. Als einzige signifikante Einflussvariable zeigt sich das GPT-Rating mit einem positiven Koeffizienten von 0,249 ($p = 0{,}018$). Steigt das GPT-Rating um eine Einheit, erhöht sich der Analystenkonsens im Durchschnitt um etwa 0,249 Einheiten. Dieser Koeffizient ist im Vergleich zum analogen Koeffizienten im GPT-Modell (0,715) deutlich kleiner, zeigt aber wieder die Gleichläufigkeit beider Bewertungsseiten. Das Kurs-Gewinn-Verhältnis weist mit einem Koeffizienten von 0,00178 ($p = 0{,}006$) einen positiven und auf dem 1-Prozent-Niveau signifikanten Effekt auf. Ein höheres KGV geht mit einem höheren, also pessimistischeren Analystenkonsens einher. Alle übrigen Variablen, einschließlich sämtlicher Branchendummies, sind in diesem Modell nicht signifikant.

Im direkten Vergleich der drei Regressionsanalysen zeigt sich, dass die Marktvolatilität des S&P 500 ausschließlich das GPT-Rating signifikant beeinflusst, während sie auf den Analystenkonsens keinen nachweisbaren Effekt hat. Bei steigender Marktvolatilität vergibt GPT also tendenziell optimistischere Ratings, die Analysten verändern ihr Urteil dagegen nicht systematisch. Umgekehrt reagiert der Analystenkonsens signifikant auf das KGV, während dieser Effekt beim GPT-Rating nicht nachweisbar ist. Die Branchenzugehörigkeit schlägt sich ebenfalls asymmetrisch nieder. Während GPT in den Sektoren Healthcare, Energy/Utilities und Automotive/Industrial Goods systematisch pessimistischere Ratings als im Technologiesektor vergibt, zeigen die Analysten kein entsprechendes branchenspezifisches Muster, der Brancheneffekt im ersten Differenzmodell geht damit primär auf die GPT-Seite zurück. Das Differenzmodell selbst bestätigt dieses Bild. Es ist mit einem p-Wert von 0,014 insgesamt signifikant und erklärt rund 61% der Varianz in der Ratingdifferenz, wobei die Marktvolatilität und die Branchenzugehörigkeit die einzigen Faktoren sind, die einen statistisch nachweisbaren Beitrag zur Erklärung der Abweichungen zwischen GPT und Analysten leisten. Unternehmensspezifische Merkmale wie Marktkapitalisierung, Bewertungskennzahlen, Dokumentengröße oder Anzahl der Analysten spielen hingegen in keinem der drei Modelle eine systematisch bedeutsame Rolle.

Tab. 4: Multiple lineare Regression der Konsens-Differenz

Koeffizienten	Schätzung	Standardfehler	t-Wert	p-Wert
(Konstante)	3.317e+00	1.795e+00	1.848	0.0745
Marktkapitalisierung	-3.964e-13	2.985e-13	-1.328	0.1942
Index-Gewicht	2.788e-01	1.982e-01	1.406	0.1699
Kurs-Gewinn-Verhältnis	-2.015e-03	1.007e-03	-2.002	0.0544
Kurs-Buchungswert-Verhältnis	-2.969e-03	2.420e-03	-1.227	0.2295
Unternehmens – Volatilität	3.170e-01	1.042e+00	0.304	0.7631
Unternehmens – Momentum	1.021e+00	5.873e-01	1.738	0.0925
Markt – Volatilität (S&P 500)	-3.578e+01	1.308e+01	-2.736	0.0103 *
Markt – Momentum (S&P 500)	1.687e+00	6.625e+00	0.255	0.8007
Anzahl der Analysten	3.003e-03	8.175e-03	0.367	0.7159
Dokumentengröße	1.523e-05	6.181e-05	0.246	0.8070
Industrie: Consumer Retail	1.778e-01	2.900e-01	0.613	0.5444
Industrie: Communication Online Services	2.589e-01	2.623e-01	0.987	0.3314
Industrie: Automotive Industrial Goods	5.276e-01	2.426e-01	2.175	0.0376 *
Industrie: Healthcare	4.780e-01	2.127e-01	2.247	0.0321 *
Industrie: Energy Utilities	6.223e-01	2.899e-01	2.147	0.0400 *
Industrie: Financial Services	6.619e-01	4.747e-01	1.394	0.1735
Industrie: Food Beverage Tobacco	5.006e-01	2.944e-01	1.700	0.0994
Industrie: Leisure Recreation	3.001e-02	4.879e-01	0.062	0.9514
Industrie: Business Professional Services	5.338e-01	4.978e-01	1.072	0.2921
$R^2 = 0{,}6081$ angepasstes $R^2 = 0{,}36$ $F(19, 30) = 2{,}45$; p-Wert = 0,01364				

Notiz: * $p < 0{,}05$; ** $p < 0{,}01$; *** $p < 0{,}001$

6.6.2 2. Regressionsmodell

Im zweiten Regressionsmodell wird für jedes der vier Zeitfenster (10, 30, 60 und 120 Handelstage) eine separate multiple Regression geschätzt, bei der die Differenz der Trefferquoten zwischen GPT und den Analysten als abhängige Variable dient. Dieselben Kontrollvariablen wie im ersten Modell werden verwendet, sodass ein direkter Vergleich der Ergebnisse über die Zeitfenster hinweg möglich ist. Ein positiver Koeffizient signalisiert dabei, dass GPT bei dem entsprechenden Faktor relativ zu den Analysten verlässlicher trifft, ein negativer Koeffizient das Gegenteil.

10-Tage-Fenster (t13): Das Modell erklärt mit einem multiplen R^2 von 0,307 rund 31% der Varianz der Trefferquotendifferenz, ist jedoch mit einem F-Wert von 0,700 und einem p-Wert von 0,790 insgesamt nicht statistisch signifikant. Auf Einzelkoeffizientenebene zeigt lediglich der Sektor Energy/Utilities einen auf dem 5-Prozent-Niveau signifikant negativen Koeffizienten von −0,759 (p = 0,032). In diesem Sektor trifft GPT im 10-Tage-Fenster um etwa 75,9 Prozentpunkte schlechter als die Analysten. Alle übrigen Variablen sind nicht signifikant.

30-Tage-Fenster (t33): Das Modell weist ein multiples R^2 von 0,441 auf, das adjustierte R^2 beträgt 0,087 und der Gesamt-F-Test ist mit F(19, 30) = 1,245 und einem p-Wert von 0,288 nicht signifikant. Als signifikante Einflussfaktoren zeigen sich das Kurs-Gewinn-Verhältnis mit einem positiven Koeffizienten von 0,000245 (p = 0,048), ein höheres KGV geht also mit einer leicht besseren relativen Trefferquote von GPT einher, sowie das Kurs-Buch-Verhältnis mit einem negativen Koeffizienten von −0,00604 (p = 0,043), was bedeutet, dass ein höheres KBV mit einer schlechteren relativen Trefferquote von GPT verbunden ist. Der Sektor Energy/Utilities weist erneut einen negativen Koeffizienten von −0,921 auf (p = 0,011) und ist damit auf dem 5-Prozent-Niveau signifikant. GPT trifft in diesem Sektor im 30-Tage-Fenster systematisch schlechter als die Analysten. Zusätzlich zeigen Anzahl der Analysten (−0,179, p = 0,073) und Dokumentengröße (−0,000128, p = 0,090) auf dem 10-Prozent-Niveau schwach negative Effekte.

60-Tage-Fenster (t63): Dieses Modell ist mit einem multiplen R^2 von 0,398, einem adjustierten R^2 von 0,017 und einem p-Wert von 0,445 das schwächste der vier Modelle und insgesamt nicht signifikant. Kein einziger Koeffizient erreicht das 5-Prozent-Niveau, lediglich der Sektor Energy/Utilities zeigt einen auf dem 10-Prozent-Niveau schwach signifikanten negativen Koeffizienten von −0,598 (p = 0,075), was auf eine ähnliche Tendenz wie in den

kürzeren Zeitfenstern hindeutet. Das 60-Tage-Fenster liefert damit insgesamt keine belastbaren Erkenntnisse über die Determinanten der relativen Treffsicherheit.

120-Tage-Fenster (t123): Das Modell erreicht mit einem multiplen R^2 von 0,539 die höchste Erklärungskraft der vier Zeitfenster, und der Gesamt-F-Test liegt mit einem p-Wert von 0,065 knapp an der Signifikanzschwelle. Als klar signifikante Einflussfaktoren zeigen sich das Unternehmensmomentum mit einem stark negativen Koeffizienten von −1,501 (p = 0,005), was bedeutet, dass GPT bei Unternehmen mit höherem positiven Momentum im 120-Tage-Fenster deutlich schlechter trifft als die Analysten. Darüber hinaus weist der Sektor Energy/Utilities einen stark negativen Koeffizienten von −0,866 (p = 0,001) auf dem 1-Prozent-Niveau auf. Die unternehmensspezifische Volatilität zeigt mit einem positiven Koeffizienten von 1,548 (p = 0,085) einen auf dem 10-Prozent-Niveau signifikanten Effekt. Bei höherer Unternehmensvolatilität trifft GPT im 120-Tage-Fenster tendenziell besser als die Analysten. Der Sektor Healthcare zeigt mit einem negativen Koeffizienten von −0,325 (p = 0,077) ebenfalls einen schwach signifikanten negativen Effekt.

Über alle vier Zeitfenster hinweg zeigt sich ein konsistentes Muster. Der Sektor Energy/Utilities ist in jedem Zeitfenster mit einem negativen Koeffizienten vertreten und in drei von vier Fenstern statistisch signifikant, was darauf hindeutet, dass die Analysten in diesem Sektor systematisch verlässlicher treffen als GPT. Die Gesamtsignifikanz der Modelle nimmt mit zunehmendem Zeithorizont zu, was darauf hinweist, dass die einbezogenen Kontrollvariablen die relative Treffsicherheit beider Seiten eher über längere als über kürzere Anlagehorizonte erklären können. Unternehmensspezifische Faktoren wie Marktkapitalisierung, Indexgewicht oder Dokumentengröße spielen hingegen in keinem der vier Modelle eine bedeutsame Rolle.

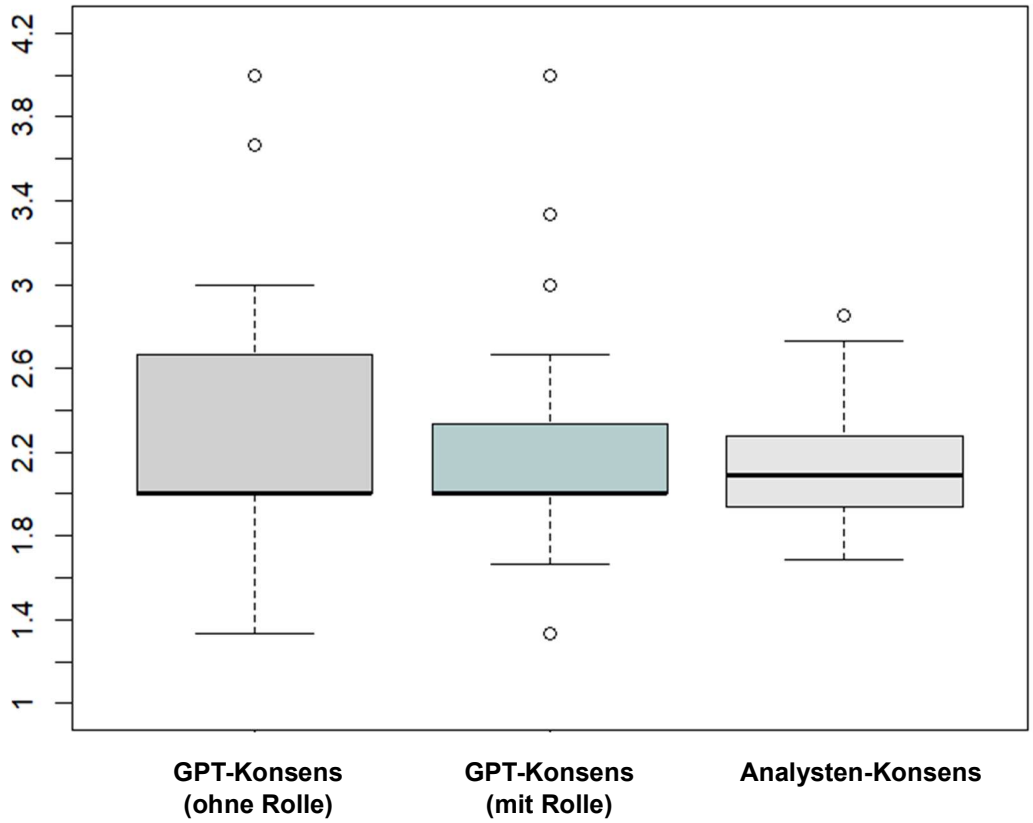
6.7 Durchlauf mit GPT – Rollenzuteilung

In Tabelle 5 wird die deskriptive Statistik des GPT-Konsenses mit Rollenvergabe neben den Daten vom GPT-Konsens ohne Rollenvergabe und dem Analystenkonsens aus Tabelle 1 dargestellt. Die Konsens-Ratings von GPT mit Rollenzuteilung reichen von 1,33 bis 4,00. Im Vergleich zu GPT_ohneRolle ist der Median von 2,00 bei GPT_mitRolle gleichgeblieben. Im Endeffekt weisen alle drei Gruppen einen Median von ca. 2,00 auf, was der Kategorie „Buy" entspricht und darauf hindeutet, dass die Mehrzahl der Empfehlungen im positiven Bereich angesiedelt sind. Der Analystenkonsens zeigt mit einem Mittelwert von 2,14 und einer Standardabweichung von 0,28 die engste Streuung aller drei Gruppen, was auf eine vergleichsweise homogene Einschätzung der Analysten hindeutet. GPT ohne Rollenangabe weist mit einem Mittelwert von 2,29 und einer Standardabweichung von 0,55 eine etwas stärkere Streuung auf, während GPT mit Rollenangabe mit einem Mittelwert von 2,23 und einer Standardabweichung von 0,48 zwischen beiden liegt. Das 3. Quartil von 2,67 bei GPT ohne Rolle gegenüber 2,33 bei GPT mit Rolle deutet darauf hin, dass die Zuweisung der Aktienanalysten-Rolle die Empfehlungen leicht in Richtung niedrigerer, also positiverer Ratingwerte verschiebt und die Streuung reduziert. In Abbildung 6 werden zur übersichtlichen Darstellung der Verteilungen aller 3 Gruppen die Boxplots nebeneinander präsentiert.

Tab. 5: Konsens-Verteilungen mit GPT-Rollenzuteilung

	1. Quartil	Median	3. Quartil	Mittelwert	Standardabweichung
GPT_ohne Rolle	2,00	2,00	2,67	2,29	0,55
Analysten-Konsens	1,94	2.09	2,28	2,14	0,28
GPT_mit Rolle	2,00	2,00	2,33	2,23	0,48

Abb. 6: Boxplots – Konsensverteilungen mit GPT-Rollenzuteilung



Zur Untersuchung, ob die explizite Zuweisung der Rolle eines professionellen Aktienanalysten die Ratings von ChatGPT verändert, wird ein gepaarter t-Test zwischen dem GPT-Konsens ohne Rollenangabe (GPT_ohneRolle) und dem GPT-Konsens mit Rollenangabe (GPT_mitRolle) durchgeführt. Da im Rahmen der Normalverteilungsprüfung ein Q-Q-Plot der Differenzen deutliche Abweichungen von der Normalverteilungslinie aufweist, insbesondere eine starke Clusterung der Differenzen im mittleren Bereich sowie ausgeprägte Abweichungen an den Rändern, wird zum t-Test zwischen GPT_mitRolle und GPT_ohneRolle ergänzend ein Wilcoxon-Vorzeichen-Rang-Test zur Robustheit der Ergebnisse berichtet.

Der gepaarte t-Test ergibt einen t-Wert von −1,565 bei 49 Freiheitsgraden und einen p-Wert von 0,124. Die mittlere Differenz zwischen GPT_mitRolle und GPT_ohneRolle beträgt −0,067, das 95-Prozent-Konfidenzintervall reicht von −0,152 bis 0,019. Der Wilcoxon-Vorzeichen-Rang-Test liefert V = 60,5 und einen p-Wert von 0,056. Beide Tests kommen damit übereinstimmend zum Ergebnis, dass kein statistisch signifikanter Unterschied zwischen

dem GPT-Konsens mit und ohne Rollenangabe nachgewiesen werden kann. Der Wilcoxon-Test verfehlt das 5-Prozent-Signifikanzniveau mit p = 0,056 jedoch nur knapp. Die negative mittlere Differenz von −0,067 zeigt, dass GPT mit Rollenangabe im Mittel leicht niedrigere, also geringfügig optimistischere Ratings vergibt als ohne Rollenangabe.

Im zweiten Vergleich wird der GPT-Konsens mit Rollenangabe dem Analystenkonsens gegenübergestellt. Der Q-Q-Plot zeigt hier eine deutlich bessere Annäherung an die Normalverteilungslinie als im ersten Vergleich, sodass ein zum t-Test ergänzender Wilcoxon-Vorzeichen-Rang-Test nicht durchgeführt wurde. Auch dieser Vergleich ergibt keinen statistisch signifikanten Unterschied (t-Wert = 1,25, p = 0,217) zwischen GPT_mitRolle und dem Analystenkonsens. Deskriptiv liegt der Mittelwert von GPT_mitRolle (2,227) zwischen dem Analystenkonsens (2,144) und GPT_ohneRolle (2,293), was einer Annäherung von GPT_mitRolle an die Analysten um 0,066 Ratingpunkte gegenüber GPT_ohneRolle entspricht.

In der Zusammenschau zeigt sich, dass keiner der zwei Vergleiche ein klar signifikantes Ergebnis auf dem 5-Prozent-Niveau liefert. Deskriptiv liegt GPT_mitRolle mit einem Mittelwert von 2,227 näher am Analystenkonsens (2,14) als GPT_ohneRolle (2,29), und auch die Standardabweichung von GPT_mitRolle (0,48) ist geringer als jene von GPT_ohneRolle (0,55), was auf eine etwas homogenere Verteilung der Ratings hindeutet, wie es beim Analystenkonsens der Fall ist. Insgesamt lässt sich festhalten, dass die Rollenangabe als professioneller Aktienanalyst die GPT-Ratings deskriptiv leicht in Richtung des Analystenkonsens verschiebt und den zuvor noch schwach sichtbaren Unterschied zu den Analysten weiter abschwächt.

Trefferquoten

Ergänzend werden die Trefferquoten von GPT mit Rollenangabe sowohl mit GPT ohne Rollenangabe als auch mit dem Analystenkonsens verglichen, um zu untersuchen, ob die explizite Zuweisung der Aktienanalysten-Rolle die Treffsicherheit von GPT beeinflusst. Die deskriptiven Befunde zeigen, dass sich die Trefferquoten aller drei Gruppen über die Zeitfenster hinweg unterschiedlich entwickeln: Im 10-Tage-Fenster erzielen GPT_mitRolle und die Analysten mit je 38% identische Trefferquoten, während GPT_ohneRolle mit 40% minimal höher liegt. Im 30-Tage-Fenster liegen alle drei Gruppen sehr nah beieinander (GPT_ohneRolle: 52%, Analysten: 52%, GPT_mitRolle: 54%). Im 60-Tage-Fenster erzielen GPT_ohneRolle und GPT_mitRolle mit je 46% identische Trefferquoten, während die Analysten mit 54% besser abschneiden. Das deutlichste Bild zeigt sich im 120-Tage-Fenster, wo die

Analysten mit einer Trefferquote von 40% sowohl GPT_ohneRolle (30%) als auch GPT_mit-Rolle (28%) klar übertreffen.

Die soeben erläuterten Trefferquoten sind in den Abbildungen 7 bis 10 visuell dargestellt. Die Fehlerbalken geben den jeweiligen Standardfehler des Mittelwerts der Trefferquote an, welcher für GPT mit Rollenzuweisung nach 10 Handelstagen 0,69, nach 30 sowie 60 Handelstagen 0,07 und nach 120 Handelstagen 0,06 beträgt. Die Standardfehler der 2 Vergleichsgruppen wurden bereits im Rahmen der Hauptanalyse zu Hypothese 2 vorgestellt.

Die gepaarten t-Tests zeigen folgendes Bild auf statistischer Ebene. Im 10-Tage-Fenster ergeben sich weder zwischen GPT_mitRolle und GPT_ohneRolle ($t = -0{,}375$, $p = 0{,}710$) noch zwischen GPT_mitRolle und den Analysten ($t = 0$, $p = 1{,}000$) signifikante Unterschiede. Gleiches gilt für das 30-Tage-Fenster, in dem beide Vergleiche keine signifikanten Differenzen aufweisen (GPT_mitRolle vs. GPT_ohneRolle: $t = 0{,}375$, $p = 0{,}710$; GPT_mitRolle vs. Analysten: $t = 0{,}330$, $p = 0{,}743$). Auch im 60-Tage-Fenster zeigen weder der Vergleich GPT_mitRolle vs. GPT_ohneRolle ($t = 0$, $p = 1{,}000$) noch der Vergleich GPT_mitRolle vs. Analysten ($t = -1{,}273$, $p = 0{,}209$) einen statistisch signifikanten Unterschied. Im 120-Tage-Fenster hingegen ergibt der Vergleich zwischen GPT_mitRolle und dem Analystenkonsens einen statistisch signifikanten Unterschied ($t = -2{,}201$, $df = 49$, $p = 0{,}032$): GPT_mitRolle trifft im Mittel um 12 Prozentpunkte schlechter als die Analysten. Der Vergleich zwischen GPT_mitRolle und GPT_ohneRolle im selben Zeitfenster bleibt hingegen nicht signifikant ($t = -0{,}444$, $p = 0{,}659$), die mittlere Differenz von −0,02 ist vernachlässigbar gering.

In der Gesamtschau legen die Trefferquoten-Vergleiche nahe, dass die Zuweisung der Aktienanalysten-Rolle die Treffsicherheit von GPT nicht verbessert. Im längsten Zeithorizont von 120 Handelstagen trifft GPT_mitRolle mit 28% sogar marginal schlechter als GPT_ohneRolle (30%) und beide GPT-Varianten liegen hier signifikant bzw. deskriptiv deutlich unter der Trefferquote der Analysten (40%). In den kürzeren Zeitfenstern von 10, 30 und 60 Handelstagen sind die Unterschiede zwischen allen drei Gruppen gering und durchgehend nicht statistisch signifikant, sodass in diesen Horizonten weder GPT_ohneRolle noch GPT_mit-Rolle gegenüber den Analysten klar unter- oder überlegen ist.

Abb. 7: Trefferquoten nach 10 Tagen (mit GPT-Rollenzuteilung)

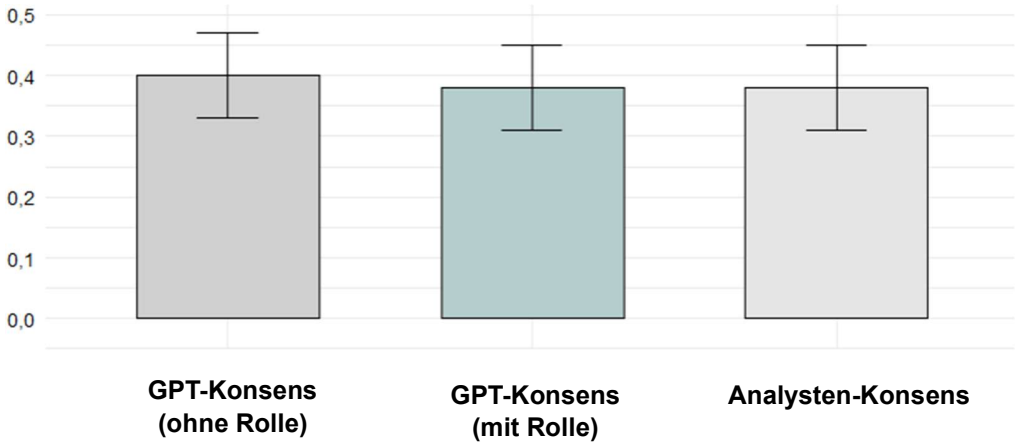


Abb. 8: Trefferquoten nach 30 Tagen (mit GPT-Rollenzuteilung)

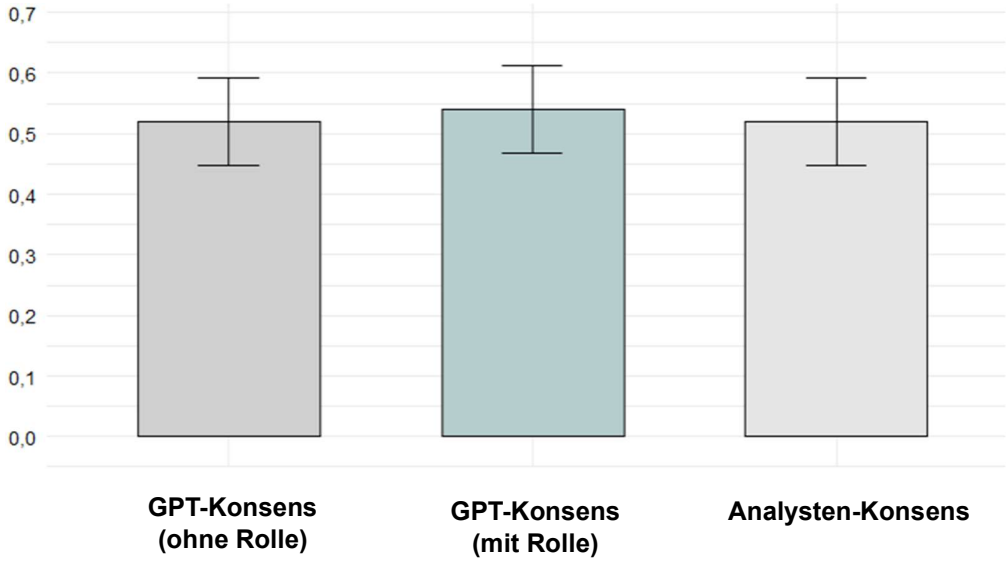


Abb. 9: Trefferquoten nach 60 Tagen (mit GPT-Rollenzuteilung)

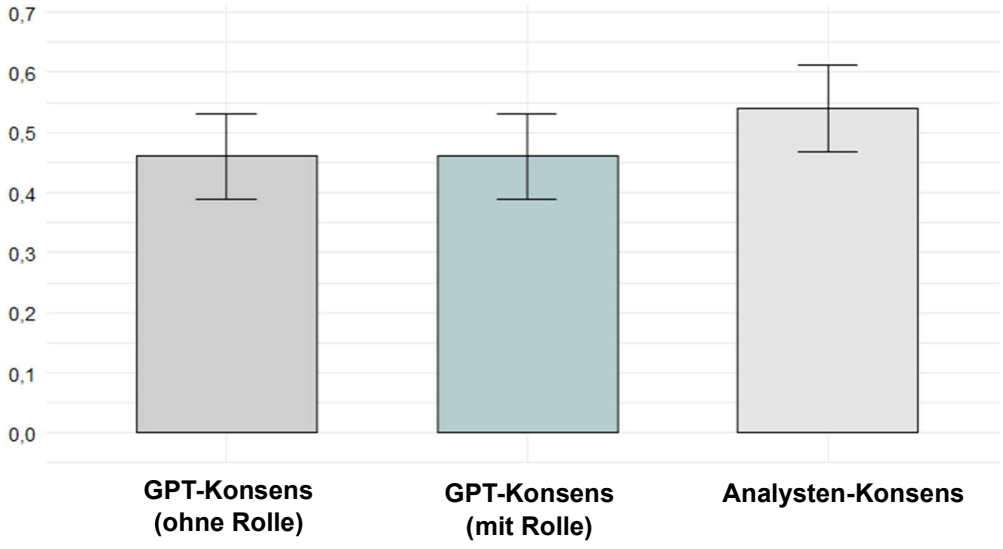
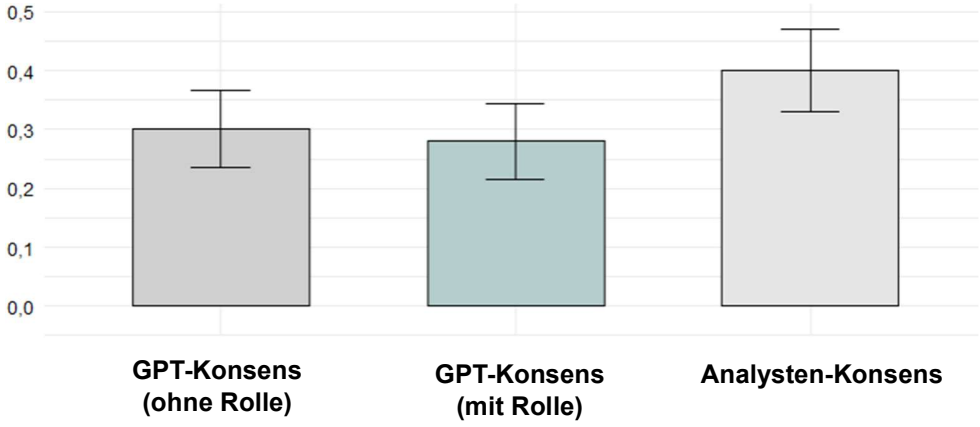


Abb. 10: Trefferquoten nach 120 Tagen (mit GPT-Rollenzuteilung)



7 Diskussion

7.1 Einordnung der Ergebnisse zu Hypothese 1

Die empirischen Befunde zu Hypothese 1 zeigen, dass sich die Handlungsempfehlungen von ChatGPT und des Analystenkonsenses auf der fünfstufigen Ratingskala im Mittel um rund 0,15 Skaleneinheiten unterscheiden, wobei GPT im Schnitt etwas pessimistischere Bewertungen vergibt als die Analysten (GPT-Mittelwert: 2,29 vs. Analysten-Mittelwert: 2,14). Der gepaarte t-Test ergibt bei einem p-Wert von 0,056 keinen statistisch signifikanten Unterschied auf dem 5-Prozent-Niveau, auf dem weniger strengen 10-Prozent-Niveau wäre er jedoch als signifikant einzustufen. Hypothese 1, dass sich die Ratings der beiden Gruppen systematisch unterscheiden, lässt sich damit auf Basis der vorliegenden Stichprobe nicht eindeutig bestätigen, liefert aber einen Hinweis auf eine tendenzielle Niveaudifferenz.

Inhaltlich ist der Befund plausibel und schließt an die Literatur an. Finanzanalysten sind bekannt dafür, systematisch optimistischere Empfehlungen zu vergeben, ein beispielsweise bei X. Li et al. (2024) thematisierter Optimismus-Bias, der unter anderem auf Interessenskonflikte, Anreizstrukturen und Reputationserwägungen zurückgeführt wird (Jegadeesh et al., 2004). ChatGPT unterliegt diesen strukturellen Anreizen in erster Linie nicht und gibt eine Empfehlung ausschließlich auf Basis der vorliegenden Berichtsinformation ab, was die leicht konservativere Tendenz der KI-Bewertungen erklären kann. Dieser Befund deckt sich auch mit den Ergebnissen von Li et al. (2024), die zeigen, dass ChatGPT-Prognosen zu Finanzkennzahlen konservativer ausfallen als jene professioneller Analysten.

Gleichzeitig deutet die substanzielle Gleichläufigkeit beider Bewertungsseiten, erkennbar am positiven Koeffizienten von 0,715 des Analystenkonsenses im GPT-Regressionsmodell darauf hin, dass GPT-5.4 Thinking bei der Verarbeitung von 10-Q-Berichten zumindest grob ähnliche Informationssignale aufgreift wie Analysten. Beide Seiten vergeben im Stichprobendurchschnitt Empfehlungen im Buy-Bereich, womit die grundsätzliche Richtung der Einschätzungen übereinstimmt. Das breitere Streuungsmaß der GPT-Empfehlungen (Standardabweichung 0,55 vs. 0,28 bei Analysten) verweist jedoch auf eine stärkere Variabilität der KI-Urteile, was möglicherweise eine Konsequenz der stochastischen Natur der Textgenerierung sowie der fehlenden institutionellen Konsensbildung und Orientierung an die Peers, die bei Analysten das Rating potentiell glätten.

7.2 Einordnung der Ergebnisse zu Hypothese 2

Hypothese 2 postuliert, dass ChatGPT weniger zuverlässige Handlungsempfehlungen liefert als Analysten, gemessen an der Übereinstimmung der empfohlenen Richtung mit der tatsächlichen späteren Kursentwicklung. Die Ergebnisse zeigen ein differenziertes Bild: In den kurzen und mittleren Prognosehorizonten von 10, 30 und 60 Handelstagen unterscheiden sich die Trefferquoten von GPT und Analysten nicht signifikant voneinander. Auf dem längsten Beobachtungshorizont von 120 Handelstagen hingegen liegen die Analysten mit einer Trefferquote von rund 40% zumindest schwach signifikant (auf dem 10%-Niveau) über jener von GPT ohne Rollenzuweisung (30%).

Hypothese 2 ist damit für den kurzfristigen, mittel- und langfristigen Bereich zu verwerfen. Auf diesen Horizonten ist keine statistisch signifikante Unterlegenheit von ChatGPT nachweisbar. Lediglich für den langfristigen Horizont von 120 Handelstagen lässt sich der deutlichste Unterscheid nachweisen. Die Analysten erweisen sich bei dieser längsten Halteperiode als schwach signifikant treffsicherer, was damit begründet werden könnte, dass die Analysten in dem Setting dieser Arbeit eine breitere Informationsbasis zur Verfügung hatten, beispielsweise auf Managementgespräche, Branchenkenntnis, historische Prognosemodelle und laufend aktualisierte Einschätzungen, die über den Inhalt des einzelnen 10-Q-Berichts hinausgehen. Je länger der Prognosehorizont, desto mehr kommen möglicherweise genau diese über den Quartalsbericht hinausgehenden Informationen zum Tragen. ChatGPT hingegen ist in diesem Studiendesign strikt auf die vorliegende Berichtsinformation beschränkt. Ein weiteres Argument für die Überlegenheit der Analysten über längere Anlagehorizonte liefert die Betrachtung ihrer Bewertungsrahmen. Aus der Untersuchung von Bradshaw (2004) geht hervor, dass Analysten ihre Empfehlungen primär auf Basis langfristiger Gewinnwachstumsprojektionen setzen, was impliziert, dass Analystenempfehlungen auf einen langfristigeren Horizont ausgerichtet sind. Dies könnte auf eine mögliche generelle Überlegenheit der Analysten gegenüber GPT in langfristigen Anlagehorizonten deuten, selbst bei gleicher Informationsbasis beider Seiten.

Bemerkenswert ist, dass die GPT-Empfehlungen im Sell-Bereich über alle Zeitfenster hinweg eine aggregierte Trefferquote von 75% aufweisen, wobei die Fallzahl von lediglich drei Beobachtungen eine valide Interpretation stark einschränkt. Die dominante Konzentration auf Buy-Empfehlungen bei beiden Gruppen, insbesondere der völlige Verzicht auf Strong-Sell-Empfehlungen, spiegelt die Asymmetrie von Analystenratings wider und deutet darauf

hin, dass auch GPT beim Lesen von Berichten börsennotierter Großunternehmen eher zu positiven Einschätzungen neigt.

Erkenntnisse zur Rollenzuteilung

Der ergänzende Durchlauf mit Rollenzuweisung zeigt, dass die explizite Instruktion an GPT, als professioneller Aktienanalyst zu agieren, die Treffsicherheit der Empfehlungen nicht verbessert. Auf keinem der drei kürzeren Prognosehorizonte ergibt sich ein signifikanter Unterschied zwischen GPT mit und ohne Rolle. Auf dem längsten Horizont von 120 Handelstagen schneidet GPT mit Rollenzuweisung mit 28% sogar marginal schlechter ab als GPT ohne Rollenzuweisung (30%) und weicht damit statistisch signifikant von der Analystentrefferquote ab ($t = -2{,}201$, $p = 0{,}032$). Dieses Ergebnis steht im Einklang mit dem Befund von Niszczota und Abbas (2023), die bei einer Rollenzuweisung als Finanzberater eine leicht schlechtere Finanzkompetenz von GPT feststellten als ohne Rollenframing. Die Rollenzuweisung scheint demnach nicht dazu beizutragen, das inhärente Wissen des Modells besser abzurufen, sondern kann die Antworten eher in eine stilistisch konditionierte Richtung verschieben, ohne den Informationsgehalt bzw. die Qualität des Outputs zu verbessern.

Einordnung in den Forschungsstand

Die vorliegenden Ergebnisse erweitern den bestehenden Forschungsstand auf mehrere Weisen. Erstens wird gezeigt, dass ChatGPT mit der aktuellen Modellgeneration (GPT-5.4 Thinking) auf kurzfristigen und mittelfristigen Horizonten mit professionellen Analysten vergleichbare Treffsicherheit bei Handlungsempfehlungen aus standardisierten Quartalsberichten erreicht. Dies ergänzt Befunde wie jene von Lopez-Lira und Tang (2025), die zeigen, dass GPT-4 aus Nachrichtenüberschriften sehr treffsichere kurzfristige Kurssignale generieren kann, nun jedoch auf den spezifischen Kontext formaler Unternehmensberichte und einen direkten Vergleich mit menschlichen Experten ausgedehnt. Zweitens bestätigt die Arbeit, dass GPT auf längeren Prognosehorizonten an Grenzen stößt (relativ zu den Analysten), wenn die Informationsbasis allein auf dem einzelnen 10-Q-Bericht beruht. Mit Berücksichtigung der Komplexität solcher Quartalsberichte, kann dies auch in Verbindung gebracht werden mit den Beobachtungen von Zhao et al. (2024) zur abnehmenden Leistungsfähigkeit von LLMs bei komplexen und langen Finanzdokumenten.

7.3 Limitationen

Die vorliegende Studie weist mehrere Einschränkungen auf, die bei der Interpretation der Ergebnisse zu berücksichtigen sind und Anknüpfungspunkte für zukünftige Forschung bieten.

Stichprobengröße und -zusammensetzung. Die Stichprobe umfasst 50 Unternehmen aus dem S&P 500, die nach Marktkapitalisierung ausgewählt wurden. Damit konzentriert sich die Analyse auf die größten und liquidesten US-Aktien, die sich durch eine besonders hohe Analystenabdeckung und öffentliche Aufmerksamkeit auszeichnen. Es ist plausibel anzunehmen, dass gerade bei diesen Unternehmen die im 10-Q-Bericht enthaltene Information bereits weitgehend im Kurs eingepreist ist, was die Treffsicherheit beider Parteien begrenzt. Eine Ausweitung der Stichprobe auf Mid- und Small-Cap-Unternehmen oder internationale Märkte könnte zu abweichenden Befunden führen.

Eingeschränkter Beobachtungszeitraum. Die Untersuchung basiert auf einem Renditeauswertungshorizont von maximal 120 Handelstagen bis Mai 2026. Ein einzelner Zeitraum kann ein spezifisches Marktregime widerspiegeln, dass die Ergebnisse verzerrt. So könnte eine Periode mit besonders starker oder schwacher Marktentwicklung oder erhöhter Volatilität die Trefferquoten beider Seiten in eine Richtung beeinflussen. Für eine robustere Aussage wären Untersuchungen über mehrere Quartale und unterschiedliche Marktphasen hinweg notwendig.

Ein längerer Beobachtungszeitraum wäre auch deshalb eine Bereicherung, weil, wie im vorherigen Abschnitt erläutert, Analysten ihre Handlungsempfehlungen eher auf einen langfristigen Anlagehorizont ausrichten. Damit könnte ein empirischer Vergleich zwischen Analysten und GPT in dem Anlagehorizont erfolgen, für den wahrscheinlich viele Analystenempfehlungen ursprünglich konzipiert sind.

Informationsasymmetrie zwischen GPT und Analysten. Das Studiendesign schränkt ChatGPT bewusst auf die Informationen des einzelnen 10-Q-Berichts ein, während Analysten typischerweise auf eine erheblich breitere Informationsbasis zurückgreifen. Diese strukturelle Asymmetrie begünstigt die Analysten, was sich insbesondere auf längeren Prognosehorizonten bemerkbar machen kann. Ein zukünftiges Design könnte GPT mit ergänzenden Informationsquellen wie aktuellen Newsdaten oder historischen Finanzzeitreihen ausstatten, um die Informationsgrundlage anzunähern und die Vergleichbarkeit zu erhöhen.

Modellvariabilität und Reproduzierbarkeit. Obwohl die stochastische Natur von GPT durch drei unabhängige Durchläufe pro Unternehmen abgefedert wurde, ist die vollständige Reproduzierbarkeit der Ergebnisse nicht gewährleistet. Modellaktualisierungen von OpenAI, Änderungen in der Sampling-Strategie oder veränderte Server-Konfigurationen können die Ausgaben beeinflussen. Zudem spiegeln die Ergebnisse ausschließlich die Performance von GPT-5.4 Thinking wider, andere Modellversionen oder Anbieter könnten zu abweichenden Befunden führen. In zukünftigen methodisch ähnlichen Untersuchungen wäre eine Aufnahme von weiteren GPT-Versionen und vor allem von anderen LLMs eine große Bereicherung, da dadurch Schlussfolgerungen nicht nur über ein Modell, sondern über die aktuelle LLM-Landschaft gezogen werden könnten.

Vereinfachung der Empfehlungskategorien. Die Aggregation der fünfstufigen Skala auf lediglich drei Richtungskategorien (positiv, neutral, negativ) und der binäre Richtigkeitsscore glätten wichtige Nuancen. Eine Buy-Empfehlung unmittelbar vor einer Kurskorrektur von −0,1% wird genauso als „falsch" gewertet wie eine Empfehlung, die einem Kursverlust von −30% vorausgeht. Differenziertere Bewertungsansätze, etwa die Berücksichtigung der Magnitude der Renditeabweichung, könnten ein präziseres Bild der relativen Prognosegüte liefern.

7.4 Fazit

Die vorliegende Arbeit untersucht, inwiefern ChatGPT auf Basis standardisierter Quartalsberichte (Form 10-Q) Handlungsempfehlungen für Aktien generieren kann, die sich qualitativ von jenen professioneller Finanzanalysten unterscheiden und deren Treffsicherheit im Hinblick auf den tatsächlichen Kursverlauf erreichen oder übertreffen. Dazu wurden für 50 S&P-500-Unternehmen die GPT-5.4-Thinking-Empfehlungen mit dem Analystenkonsens verglichen und anhand der realisierten Renditen in vier Prognosefenstern bewertet.

Zu **Hypothese 1**, dass sich die Ratings von GPT und Analysten unterscheiden, liefert die Analyse einen Hinweis auf eine tendenzielle Niveaudifferenz: GPT bewertet im Mittel etwas konservativer als die Analysten, was mit dem bekannten Optimismus-Bias professioneller Analysten konsistent ist. Auf dem konventionellen 5-Prozent-Signifikanzniveau lässt sich ein formaler statistischer Unterschied jedoch nicht nachweisen. Hypothese 1 kann damit nicht eindeutig bestätigt werden.

Zu **Hypothese 2**, dass ChatGPT weniger zuverlässige Empfehlungen liefert als Analysten, zeigen die Ergebnisse ein nuanciertes Bild. Für die kurzfristigen und mittelfristigen Horizonte von 10, 30 und 60 Handelstagen sind die Trefferquoten beider Gruppen statistisch nicht unterscheidbar und auf diesen Horizonten ist Hypothese 2 zu verwerfen. Für den längsten untersuchten Horizont von 120 Handelstagen ist Hypothese 2 auf dem 5-Prozent-Niveau ebenfalls zu verwerfen, obwohl sich Analysten als schwach signifikant (auf dem 10-Prozent-Niveau) treffsicherer erweisen.

Aus methodischer Perspektive zeigt die Studie, dass die Zuweisung einer Analysten-Rolle an GPT die Treffsicherheit nicht steigert und auf dem längsten Horizont sogar marginal mindert, was sogar zu einer statistisch signifikanten (auf dem 5-Prozent-Niveau) Überlegung der Analysten im längsten Beobachtungszeitraum führt. Dies lässt darauf deuten, dass Verbesserungen der GPT-Performance im Finanzkontext weniger durch Rollenframing als durch eine reichhaltigere Informationsbasis oder spezialisiertes Finetuning erzielt werden dürften.

Insgesamt legen die Ergebnisse nahe, dass GPT-5.4 Thinking bei der Verarbeitung formaler Unternehmensberichte auf kurzfristigen Horizonten ein beachtliches Maß an Prognosekompetenz erreicht und damit als unterstützendes Instrument in der Finanzanalyse Potenzial besitzt. Auf längeren Horizonten bleibt der Mensch, zumindest in der Form des kollektiven Analystenkonsenses mit seiner breiteren Informationsgrundlage, überlegen. Mit dem Fortschreiten der Modellgenerationsentwicklung, der Erweiterung der Kontextfenster und der Integration von Echtzeit-Datenquellen für die KI könnte sich dieses Bild jedoch verschieben. Zukünftige Forschung sollte das hier entwickelte Design auf mehrere Quartale ausdehnen, die Informationsbasis von ChatGPT systematisch erweitern und unterschiedliche Modelle sowie alternative Märkte vergleichen, um robustere und generalisierbarere Aussagen zur Leistungsfähigkeit von LLMs im Bereich der Aktienempfehlung treffen zu können.

Literaturverzeichnis

Aleksandrova, A., Ninova, V., & Zhelev, Z. (2023). A Survey on AI Implementation in Finance, (Cyber) Insurance and Financial Controlling. *Risks*, *11*(5), 91. https://doi.org/10.3390/risks11050091

Balsiger, D., Dimmler, H.-R., Egger-Horstmann, S., & Hanne, T. (2024). Assessing Large Language Models Used for Extracting Table Information from Annual Financial Reports. *Computers*, *13*(10), 257. https://doi.org/10.3390/computers13100257

Barman, D., Guo, Z., & Conlan, O. (2024). The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination. *Machine Learning with Applications*, *16*, 100545. https://doi.org/10.1016/j.mlwa.2024.100545

Beccalli, E., Miller, P., & O'leary, T. (2015). How Analysts Process Information: Technical and Financial Disclosures in the Microprocessor Industry. *European Accounting Review*, *24*(3), 519–549. https://doi.org/10.1080/09638180.2014.969526

Bilinski, P. (2024). The Usefulness of ChatGPT for Textual Analysis of Annual Reports. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4723503

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2023). Textual Analysis in Accounting: What's Next? *Contemporary Accounting Research*, *40*(2), 765–805. https://doi.org/10.1111/1911-3846.12825

Boyson, N., Cao, Y., Liu, M., & Wan, C. (2023). *EXECUTIVES VS CHATBOTS: UNMASKING INSIGHTS THROUGH HUMAN-AI DIFFERENCES IN EARNINGS CONFERENCE Q&A*.

Bradshaw, M. T. (2004). How Do Analysts Use Their Earnings Forecasts in Generating Stock Recommendations? *The Accounting Review*, *79*(1), 25–50.

Buergler, V., Davidson, B., & Bürgler, M. (2025). *Democratizing Financial Information using GPT-4o*. SSRN. https://doi.org/10.2139/ssrn.5130935

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, *6*(6), 101260. https://doi.org/10.1016/j.patter.2025.101260

Cheng, X., Nanjundan, P., & George, J. P. (2025). *Introduction to Natural Language Processing: Exploring Techniques, Applications, and Challenges*. River Publishers. https://doi.org/10.1201/9781003660675

Cheng, Y., Zeng, Y., & Zou, J. (2026). Harnessing ChatGPT for predictive financial factor generation: A new frontier in financial analysis and forecasting. *The British Accounting Review*, *58*(2), 101507. https://doi.org/10.1016/j.bar.2024.101507

Desai, H., Liang, B., & Singh, A. K. (2000). Do All-Stars Shine? Evaluation of Analyst Recommendations. *Financial Analysts Journal*, *56*(3), 20–29. https://doi.org/10.2469/faj.v56.n3.2357

Dong, M. M., Stratopoulos, T. C., & Wang, V. X. (2024). A scoping review of ChatGPT research in accounting and finance. *International Journal of Accounting Information Systems*, *55*, 100715. https://doi.org/10.1016/j.accinf.2024.100715

Hao, Y. (2025). Large Language Models in Accounting and Financial Research: A Review of Applications in Accounting-related Text Analysis. *Advances in Economics, Management and Political Sciences*, *189*, 55–60. https://doi.org/10.54254/2754-1169/2025.BL24036

Harris, T. (2025). Managers' perception of product market competition and earnings management: A textual analysis of firms' 10-K reports. *Journal of Accounting Literature*, *47*(3), 499–524. https://doi.org/10.1108/JAL-11-2022-0116

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, *241*, 122666. https://doi.org/10.1016/j.eswa.2023.122666

Jegadeesh, N., Kim, J., Krische, S. D., & Lee, C. M. C. (2004). Analyzing the Analysts: When Do Recommendations Add Value? *The Journal of Finance*, *59*(3), 1083–1124. https://doi.org/10.1111/j.1540-6261.2004.00657.x

Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*.

Kadam, S., & Sethi, M. (2024). Target price accuracy of sell-side analysts: Evidence from India. *Cogent Economics & Finance*, *12*(1), 2423261. https://doi.org/10.1080/23322039.2024.2423261

Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *Finance Research Letters*, *62*, 105227. https://doi.org/10.1016/j.frl.2024.105227

Ko, H., & Lee, J. (2024). Can ChatGPT improve investment decisions? From a portfolio management perspective. *Finance Research Letters*, *64*, 105433. https://doi.org/10.1016/j.frl.2024.105433

Kovorova-Simacek, M., Zülch, H., Kirschbaum, L., Klammer, K., Barrantes, E., Horvathova, A., & Schilling, C. (2026). *GenAI as Reader: How ChatGPT & Co. Use Annual Reports*. USTP – University of Applied Sciences St. Pölten, HHL Leipzig Graduate School of Management & nexxar. https://digital-investor-relations.com/_assets/downloads/DIR_GenAI-as-Reader.pdf?h=E0wP6LL9

Kuroki, Y., Manabe, T., & Nakagawa, K. (2023). *Fact or Opinion? – Essential Value for Financial Results Briefing* (SSRN Scholarly Paper Nr. 4430511). Social Science Research Network. https://doi.org/10.2139/ssrn.4430511

Li, D., Li, Y., Chen, X., & Zhang, W. (2026). Augmenting large language models for financial sentiment analysis: A heuristic sparse mixture-of-experts framework. *PeerJ Computer Science*, *12*, e3607. https://doi.org/10.7717/peerj-cs.3607

Li, H., Chen, X., XU, Z., Li, D., Hu, N., Teng, F., Li, Y., Qiu, L., Zhang, C. J., Li, Q., & Chen, L. (2025). *Exposing Numeracy Gaps: A Benchmark to Evaluate Fundamental Numerical Abilities in Large Language Models* (arXiv:2502.11075). arXiv. https://doi.org/10.48550/arXiv.2502.11075

Li, J., Li, G., Li, Y., & Jin, Z. (2025). Structured Chain-of-Thought Prompting for Code Generation. *ACM Trans. Softw. Eng. Methodol.*, *34*(2), 37:1-37:23. https://doi.org/10.1145/3690635

Li, W., Liu, W., Deng, M., Liu, X., & Feng, L. (2025). The impact of large language models on accounting and future application scenarios. *Journal of Accounting Literature*. https://doi.org/10.1108/JAL-12-2024-0357

Li, X., Feng, H., Yang, H., & Huang, J. (2024). Can ChatGPT reduce human financial analysts' optimistic biases? *Economic and Political Studies*, *12*(1), 20–33. https://doi.org/10.1080/20954816.2023.2276965

Li, Y., Alexander, N., & Hong, T. S. (2020). Opportunity knocks but once: Delayed disclosure of financial items in earnings announcements and neglect of earnings news. *Review of Accounting Studies*, *25*(1), 159–200. https://doi.org/10.1007/s11142-019-09519-7

Lopez-Lira, A., & Tang, Y. (2025). *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models* (arXiv:2304.07619). arXiv. https://doi.org/10.48550/arXiv.2304.07619

Magner, N., Henríquez, P. A., & Sanhueza, A. (2025). Decoding risk sentiment in 10-K filings: Predictability for U.S. stock indices. *Finance Research Letters*, *81*, 107472. https://doi.org/10.1016/j.frl.2025.107472

Niszczota, P., & Abbas, S. (2023). GPT has become financially literate: Insights from financial literacy tests of GPT and a preliminary test of how people use it as a source of advice. *Finance Research Letters*, *58*, 104333. https://doi.org/10.1016/j.frl.2023.104333

OpenAI. (o. J.-a). *Model Release Notes*. OpenAI Help Center. Abgerufen 9. Mai 2026, von https://help.openai.com/en/articles/9624314-model-release-notes

OpenAI. (2022). *ChatGPT ist da*. https://openai.com/de-DE/index/chatgpt/

OpenAI, A. (o. J.-b). *GPT-5.4, OpenAI Developers*. Abgerufen 9. Mai 2026, von https://developers.openai.com/api/docs/models/gpt-5.4

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.02155

Pelster, M., & Val, J. (2024). Can ChatGPT assist in picking stocks? *Finance Research Letters*, *59*, 104786. https://doi.org/10.1016/j.frl.2023.104786

Qian, J., Wang, H., Li, Z., Li, S., & Yan, X. (2022). *Limitations of Language Models in Arithmetic and Symbolic Induction* (arXiv:2208.05051). arXiv. https://doi.org/10.48550/arXiv.2208.05051

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

Schneider, C. J., & Yilmaz, Y. (2025). Stock portfolio selection based on risk appetite: Evidence from ChatGPT. *Finance Research Letters*, *82*, 107517. https://doi.org/10.1016/j.frl.2025.107517

SEC. (o. J.). *Form 10-Q*. Abgerufen 1. Mai 2026, von https://www.sec.gov/files/form10-q.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Welsch, A., Eitle, V., & Buxmann, P. (2018). Maschinelles Lernen. *HMD Praxis der Wirtschaftsinformatik*, *55*(2), 366–382. https://doi.org/10.1365/s40702-018-0404-z

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT* (arXiv:2302.11382). arXiv. https://doi.org/10.48550/arXiv.2302.11382

Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Suchow, J. W., Zhang, D., & Khashanah, K. (2025). FinMem: A Performance-Enhanced LLM Trading Agent With Layered Memory and Character Design. *IEEE Transactions on Big Data*, *11*(6), 3443–3459. https://doi.org/10.1109/TBDATA.2025.3593370

Zhao, Y., Long, Y., Liu, H., Kamoi, R., Nan, L., Chen, L., Liu, Y., Tang, X., Zhang, R., & Cohan, A. (2024). DocMath-Eval: Evaluating Math Reasoning Capabilities of LLMs in Understanding Long and Specialized Documents. In L.-W. Ku, A. Martins, & V. Srikumar (Hrsg.), *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers) (S. 16103–16120). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.852

Hinweis zur Verwendung von Künstlicher Intelligenz

Für die sprachliche Optimierung, Grammatikprüfung und Stilverbesserung dieser Masterarbeit wurde das Large Language Model „Claude Sonnet 4.6“ verwendet. Die inhaltliche Ausarbeitung, Strukturierung sowie die wissenschaftlichen Schlussfolgerungen wurden vollständig und eigenständig vom Autor verfasst.