Analyse von Sprachmodellen mit Methoden der experimentellen Wirtschaftsforschung

Seminararbeit

Alfred-Weber-Institut für Wirtschaftswissenschaften (AWI)
Universität Heidelberg
Wintersemester 2025/26

Dozent: Patrick Wolfgang Schmidt
Studiengang: B.Sc. Volkswirtschaftslehre
Name: Moritz Graemer
Matrikelnummer: 4734368

Heidelberg, 14.01.2026

Inhaltsverzeichnis

Analyse von Sprachmodellen: Overconfidence in Trading Situations

Moritz Graemer

14.01.2026

1 Einleitung

Finanzmärkte sind durch Unsicherheit, hohe Informationsdichte und komplexe Entscheidungsumgebungen gekennzeichnet. In solchen Kontexten spielt nicht nur die Qualität von Prognosen eine zentrale Rolle, sondern auch die Einschätzung der eigenen Sicherheit. Die verhaltensökonomische Literatur zeigt, dass menschliche Akteure systematisch von Overconfidence betroffen sind, insbesondere in Form von Overprecision: einer Überschätzung der Wahrscheinlichkeit, dass die eigene Einschätzung korrekt ist. Diese Verzerrung ist eng mit suboptimalem Entscheidungsverhalten auf Finanzmärkten verknüpft und äußert sich unter anderem in Fehlkalibrierung von Wahrscheinlichkeiten, übermäßigem Vertrauen in Signale und einer Unterschätzung von Unsicherheit (Barber and Odean (2001), Glaser and Weber (2004), Karki et al. (2024)). Mit dem zunehmenden Einsatz von Large Language Models (LLMs) in ökonomischen Entscheidungsprozessen, etwa als Analysewerkzeuge, Prognosehilfen oder potenzielle Finanzberater, stellt sich die Frage, ob und in welcher Form diese Modelle ähnliche systematische Verzerrungen aufweisen. Während LLMs in vielen Aufgaben eine hohe prognostische Leistungsfähigkeit zeigen, ist bislang wenig darüber bekannt, wie gut ihre subjektiven Wahrscheinlichkeitsaussagen kalibriert sind und wie sie auf unterschiedliche Informationsumgebungen reagieren. Insbesondere ist unklar, ob sich zusätzliche Information und ihr tatsächlichen Prognosegehalt, zu einer unbegründeten Erhöhung der Konfidenz führt und damit Overconfidence erzeugt. Ziel dieser Arbeit ist es, das Entscheidungsverhalten von LLMs im Kontext täglicher Investmententscheidungen systematisch zu untersuchen. Dazu wird ein kontrolliertes, sequentielles Entscheidungssetting implementiert, in dem Modelle täglich eine binäre Investitionsentscheidung treffen und gleichzeitig eine subjektive Wahrscheinlichkeit für die Richtung der Marktentwicklung angeben. Der Fokus liegt dabei auf der Beziehung zwischen realisiertem Outcome und subjektiver Sicherheit. Die zentrale Forschungsfrage dieser Arbeit lautet:

Wie beeinflussen unterschiedliche Informationsumgebungen die Kalibration, Overconfidence und Gesamtgüte probabilistischer Entscheidungen von Large Language Models im Finanzkontext?

Zur Beantwortung dieser Frage werden drei Hypothesen formuliert:

- **H1 (Baseline-Kalibration):** Ohne zusätzliche Information geben LLMs Wahrscheinlichkeiten nahe der Basisrate an und zeigen eine gute Kalibration mit geringer Resolution.
- **H2 (Irrelevante Information und Overprecision):** Irrelevante, nicht-diagnostische Informationen erhöhen die subjektive Konfidenz der Modelle, ohne die Prognosequalität zu verbessern, und führen zu höherer Fehlkalibrierung (Overprecision).
- **H3 (Relevante Information und Lernübertreibung):** Relevante Informationen verbessern die Gesamtgüte der Vorhersagen durch höhere Resolution, gehen jedoch mit einer überproportionalen Zunahme der Konfidenz einher, sodass die Kalibration schlechter wird.

Methodisch wird diese Fragestellung mithilfe des Brier Scores und des Expected Calibration Error (ECE) analysiert, die eine Trennung zwischen Gesamtgüte, Trennschärfe und Kalibration erlauben. Durch den Vergleich mehrerer State-of-the-art-Modelle wird zudem untersucht, ob sich diese Effekte modellabhängig unterscheiden oder ein generelles Muster im Verhalten von LLMs darstellen.

2 Literatur

Overconfidence wird in der Literatur in drei Unterkategorien unterteilt: **Overestimation**, **Overplacement** und **Overprecision**. Während Overestimation die Überschätzung der eigenen absoluten Fähigkeiten beschreibt und Overplacement die Überschätzung der eigenen Fähigkeiten relativ zu anderen, bezeichnet Overprecision eine systematische Überschätzung der Wahrscheinlichkeit, dass die eigene Einschätzung korrekt ist (Moore and Healy (2008); Prims and Moore (2017)).

Zentrale theoretische Arbeiten zeigen zudem, dass Overconfidence stark von der Aufgabenstruktur abhängt: Bei einfachen Aufgaben tritt häufig Underconfidence auf, während bei komplexen und unsicheren Entscheidungsproblemen Overconfidence dominiert (Moore and Healy (2008)).

Im finanzökonomischen Kontext ist Overconfidence eng mit suboptimalem Entscheidungsverhalten verknüpft. Eine systematische Literaturübersicht zeigt, dass Overconfidence Investitionsentscheidungen verzerrt und zu ineffizientem Verhalten beiträgt (Karki et al. (2024)). Empirische Evidenz legt nahe, dass dieser Bias insbesondere bei männlichen

Investoren stärker ausgeprägt ist und sich unter anderem in höherer Handelsaktivität äußert (Barber and Odean (2001)). Gleichzeitig zeigen andere Studien, dass Overconfidence nicht zwingend mit Handelsvolumen korreliert ist, insbesondere wenn Overconfidence über Fehlkalibrierung und nicht über Selbstüberschätzung gemessen wird (Glaser and Weber (2004)).

Darüber hinaus zeigt die Literatur, dass sowohl unerfahrene als auch professionelle Marktteilnehmer systematisch overprecise sind. Zwar passen erfahrene Akteure ihre Einschätzungen nach Erfolgen und Misserfolgen in bayesianischer Weise an, diese Anpassung ist jedoch unzureichend, um vollständige Kalibration zu erreichen (Merkle and Schreiber (2025)). Overconfidence, insbesondere Overprecision, kann dabei auf unterschiedliche Weise gemessen werden. Ein verbreiteter Ansatz besteht darin, Entscheidungsträger explizit nach Konfidenzintervallen oder subjektiven Wahrscheinlichkeiten zu fragen (Binnendyk and Pennycook (2024)). Gleichzeitig wird darauf hingewiesen, dass das Ausmaß beobachteter Overconfidence stark vom Messdesign abhängt und insbesondere durch die explizite Abfrage von Unsicherheit verstärkt werden kann (Klayman et al. (1999)).

Neuere Arbeiten übertragen diese Konzepte auf LLMs. Erste Befunde zeigen, dass LLMs menschliche Muster von Overconfidence teilweise spiegeln, jedoch weniger sensitiv auf Aufgabenschwierigkeit reagieren. Stattdessen adaptieren sie stereotypisches Entscheidungsverhalten, wenn ihnen entsprechende Rollen oder Narrative im Prompt zugewiesen werden, selbst wenn sich die objektive Genauigkeit nicht verändert (Xu et al. (2025)). In hochrelevanten und komplexen Entscheidungssituationen, etwa wenn LLMs als Richter eingesetzt werden, tritt Overconfidence ebenfalls systematisch auf (Tian et al. (2026)). Darüber hinaus zeigen aktuelle Studien, dass LLMs insbesondere dann stark überkonfident werden, wenn sie unsicher sind, und dabei bestehende menschliche Biases verstärken können (Sun et al. (2025)). Da Overconfidence zudem mit höherer Verschuldung und riskanteren finanziellen Entscheidungen korreliert ist, ergeben sich daraus relevante Implikationen für den Einsatz von LLMs als potenzielle Finanzberater (Grohmann et al. (2023)).

Vor diesem Hintergrund fokussiert sich die vorliegende Arbeit auf Overprecision im Finanzkontext und misst diese explizit über von LLMs angegebene subjektive Wahrscheinlichkeiten. Durch den Vergleich vorhergesagter Konfidenzen mit realisierten Marktereignissen wird untersucht, ob und wie unterschiedliche Informationsumgebungen systematische Fehlkalibrierung erzeugen.

3 Ablauf des Experiments

Das vorliegende Experiment implementiert ein sequentielles Entscheidungssetting, um Overconfidence und Kalibrierung von LLMs im Kontext täglicher Investmententscheidungen zu untersuchen. Konkret wird das Modell in die Rolle eines "rationalen Inves-

tors"versetzt, der an jedem Handelstag eine binäre Entscheidung trifft (buy oder sell) und zusätzlich eine subjektive Wahrscheinlichkeit dafür angibt, dass der DAX höher schließt als am Morgen. Für jeden Tag erhält das Modell neben dem aktuellen Datum und dem heutigen DAX-Eröffnungskurs auch die vollständige Historie seiner eigenen bisherigen Entscheidungen in diesem Run.

Für jeden Tag werden aus dem Eröffnungs- (open) und Schlusskurs (close) ein binärer Outcome (gain) konstruiert, der den realisierten Marktzustand abbildet: gain ist 1, wenn der Schlusskurs über dem Eröffnungskurs liegt, sonst 0 . Diese Reduktion auf eine binäre Zielvariable ist eine bewusste Modellannahme. Sie abstrahiert von Renditehöhen, Volatilität oder Dynamiken zwischen Marktöffnungszeiten und fokussiert ausschließlich auf die Richtung der Tagesbewegung. Dadurch wird klar definiert, was als richtig oder falsch gilt. Der genaue Prompt ist in der Appendix zu finden. A

Die empirische Grundlage bilden reale DAX-Daten, die über `yfinance` für einen Zeitraum von 30 Kalendertagen heruntergeladen werden. Dies führt zu einer Datenmenge von 16 Handelstagen. Die geringe Anzahl ist durch die hohe Rechenleistung geschuldet, im Programm aber mit minimalem Aufwand zu erweitern. Die Daten werden möglichst aktuell gewählt, um zu vermeiden, dass das LLM die genauen Kursdaten bereits gelernt hat und somit keine Unsicherheit über die Entwicklung des DAX besteht. Falls die Annahme in Frage gestellt wird, dass das Modell die Kurse nicht schon gelernt hat, oder sogar im Internet danach gesucht hat, könnte man die Kurse mit einem Faktor verzerren, um sie für das Modell unkenntlich zu machen.

Ein weiterer Kernbestandteil des Designs sind die drei Treatments: "without information", "with irrelevant information", "with relevant information". Im Baseline-Treatment erhält das Modell ausschließlich den DAX-Eröffnungskurs, das Datum und die eigene Entscheidungshistorie. Damit wird ein extrem informationsarmes Umfeld geschaffen, in dem das Modell objektiv kaum nützliche Signale zur Entwicklung des Kurses zur Verfügung hat. Jede systematische Abweichung von gut kalibrierter Unsicherheit, also Konfidenzen weit weg von 0.5, kann hier als Ausdruck eingeschlossener Modellbiases interpretiert werden.

Das Treatment mit irrelevanten Informationen ergänzt den Prompt um faktisch korrekte, aber für die kurzfristige Tagesrichtung ökonomisch bedeutungslose Aussagen über den DAX (z. B. "The name DAX stands for German Stock Index and is a registered trademark."). Die Annahme hier ist, dass solche Informationen keinen rationalen Informationsgehalt für die Prognose haben, aber kognitiv als Information wahrgenommen werden können. Wenn sich in diesem Treatment eine systematisch höhere Konfidenz oder schlechtere Kalibrierung zeigt als im Baseline-Fall, deutet dies auf eine Form von Overconfidence durch irrelevante Signale hin. Dies wäre analog zu menschlichem Verhalten, bei dem zusätzliche, aber nutzlose Informationen das subjektive Sicherheitsgefühl erhöhen.

Das Treatment mit relevanten Informationen fügt dagegen makroökonomische, tech-

nische und sentimentbezogene Aussagen hinzu, die plausibel mit kurzfristigen Marktbewegungen zusammenhängen könnten, ohne jedoch konkret auf den jeweiligen Handelstag zugeschnitten zu sein (z.B. Technical analysis shows the DAX attempting to break above long-term resistance levels, suggesting traders are watching key psychological price zones."). Die Informationen sind absichtlich unscharf, allgemein und nicht quantifiziert. Sie simulieren typische Finanznachrichten oder Marktkommentare, wie sie menschliche Investoren konsumieren. Damit wird getestet, ob mehr relevante Information tatsächlich zu besserer Kalibrierung führt, oder ob sie lediglich die Konfidenz erhöht, ohne die Prognosequalität zu verbessern.

Die Situation wird mit den vier typischsten State-of-the-art Modellen analysiert: Claude Opus, Gemini Flash , Llama-4-Maverick und GPT-5-mini. Methodisch wird jedes (Modell $\times$ Treatment)-Setting drei Mal wiederholt, um stochastische Effekte durch Sampling (Temperatur = 1.0) zu erfassen. Jeder Run ist eine vollständige Sequenz über alle Tage, und die Wiederholungen erlauben es, Mittelwerte von Brier Score, ECE und Entscheidungsmetriken zu berechnen. Die beobachtete Varianz zwischen Runs geht also nur auf die modelleigene Stochastik des LLM zurück und nicht auf Marktunterschiede.

4 Analysemethoden

Für die Analyse werden zwei Maße verwendet: Der Brier Score und der Expected Calibration Error. Ergänzend werden einfache Entscheidungsmetriken berechnet, etwa die Anzahl der Buy-Tage oder der korrekt getroffenen Investmententscheidungen. Diese zusätzlichen Kennzahlen beruhen auf der Annahme, dass das Modell tatsächlich eine implizite Entscheidungsregel verfolgt (z. B. buy bei Konfidenz ≥ 0.5), auch wenn diese Regel nicht explizit erzwungen wird. Sie dienen weniger der Performancebewertung im ökonomischen Sinne als der Kontextualisierung der Kalibrierungsmaße. Die vom Modellen ausgegebenen Wahrscheinlichkeiten und Entscheidungen werden analysiert, indem ihre probabilistischen Vorhersagen mit den realisierten binären Marktereignissen (gain $Y = 1$/ loss $Y = 0$) vergleichen werden. Das optimale Modell hat eine gute Kalibrierung und eine gute Gesamtgüte. Kalibrierung: Wenn ein Modell häufig confidence = 0.7 ausgibt, sollte das Ereignis in etwa 70% dieser Fälle eintreten ($P(Y = 1) = 0.7$). Wenn man zu der Kalibrierung die Unsicherheit des Marktes und die Fallunterscheidungen des Modells hinzunimmt, erhält man die Gesamtgüte. Diese werden durch Brier Score und Expected Calibration Error (ECE) gemessen.

4.1 Brier Score

Der Brier Score misst den mittleren quadratischen Fehler zwischen vorhergesagter Wahrscheinlichkeit und realisiertem Outcome und bestraft sowohl Über- als auch Unterkonfi-

denz.

$$Brier = \frac{1}{N}\sum_{t=1}^{N}(p_t - y_t)^2$$

Eine sehr selbstsichere, aber falsche Prognose (p nahe 1 bei y=0) wird deutlich stärker bestraft als eine vorsichtige Prognose nahe 0.5. Somit kann Overconfidenz als Abweichung der geschätzten Wahrscheinlichkeit von der Realität mit dem Brier Score geschätzt werden. Durch das einfache Design wird gewährleistet, dass Overconfidence nicht als Artefakt komplexer Marktmechanik, sondern als Eigenschaft des Entscheidungsprozesses des Modells selbst auftritt.

Ein wichtiger Vorteil des Brier Scores ist seine Zerlegung in Uncertainty, Resolution und Reliability B:

$$\text{Brier} = \underbrace{\mathbb{V}(Y)}_{\text{Uncertainty}} - \underbrace{\mathbb{V}(\mathbb{E}[Y|\hat{p}])}_{\text{Resolution}} + \underbrace{\mathbb{E}[(\hat{p} - \mathbb{E}[Y|\hat{p}])^2]}_{\text{Reliability}}$$

- **Uncertainty** ist Markt-Rauschen
- **Resolution** misst, ob das Modell überhaupt zwischen Situationen mit höherer und niedrigerer Eintrittswahrscheinlichkeit differenziert. Wenn ein Modell in manchen Fällen eher 0.6 und in anderen eher 0.4 prognostiziert und diese Unterschiede mit realen Häufigkeiten korrespondieren, steigt die Resolution.
- **Reliability / Kalibrierung** entspricht Fehlkalibrierung: Sie steigt, wenn die ausgegebenen Wahrscheinlichkeiten systematisch vom wahren bedingten Eintrittsanteil abweichen.

Im Baseline- und Irrelevanz-Treatment ist geringe Resolution zu erwarten, weil keine echten Signale geliefert werden, während das relevante Treatment zumindest theoretisch mehr Resolution ermöglichen sollte. Gleichzeitig kann die Reliability gerade im Irrelevanz-Treatment schlechter werden, wenn irrelevanter Kontext die Konfidenz inflationiert.

4.2 Expected Calibration Error (ECE)

Der ECE misst explizit die Kalibrierung (Reliability). Er aggregiert Vorhersagen in Bins über den Wahrscheinlichkeitsraum (im Programm 10 Bins in [0,1]) und vergleicht pro Bin die mittlere vorhergesagte Wahrscheinlichkeit mit der empirischen Trefferquote:

$$ECE = \sum_{b=1}^{B}\frac{n_b}{N}|\hat{p}_b - \hat{y}_b|$$

- n_b die Anzahl der Beobachtungen im Bin b

- $\hat{p}_b$ die durchschnittliche prognostizierte Wahrscheinlichkeit in b
- $\hat{y}_b$ die beobachtete Häufigkeit von Y=1 in b

ECE = 0 bedeutet perfekte Kalibrierung. Höhere Werte zeigen systematische Over- oder Underconfidence. Wenn irrelevante Informationen ein Sicherheitsgefühl erzeugen, sollte das Modell häufiger hohe p-Werte ausgeben, ohne dass die empirische Trefferquote $\hat{y}_b$ in diesen Bereichen entsprechend steigt. Der ECE würde somit steigen.

5 Analyse der Ergebnisse

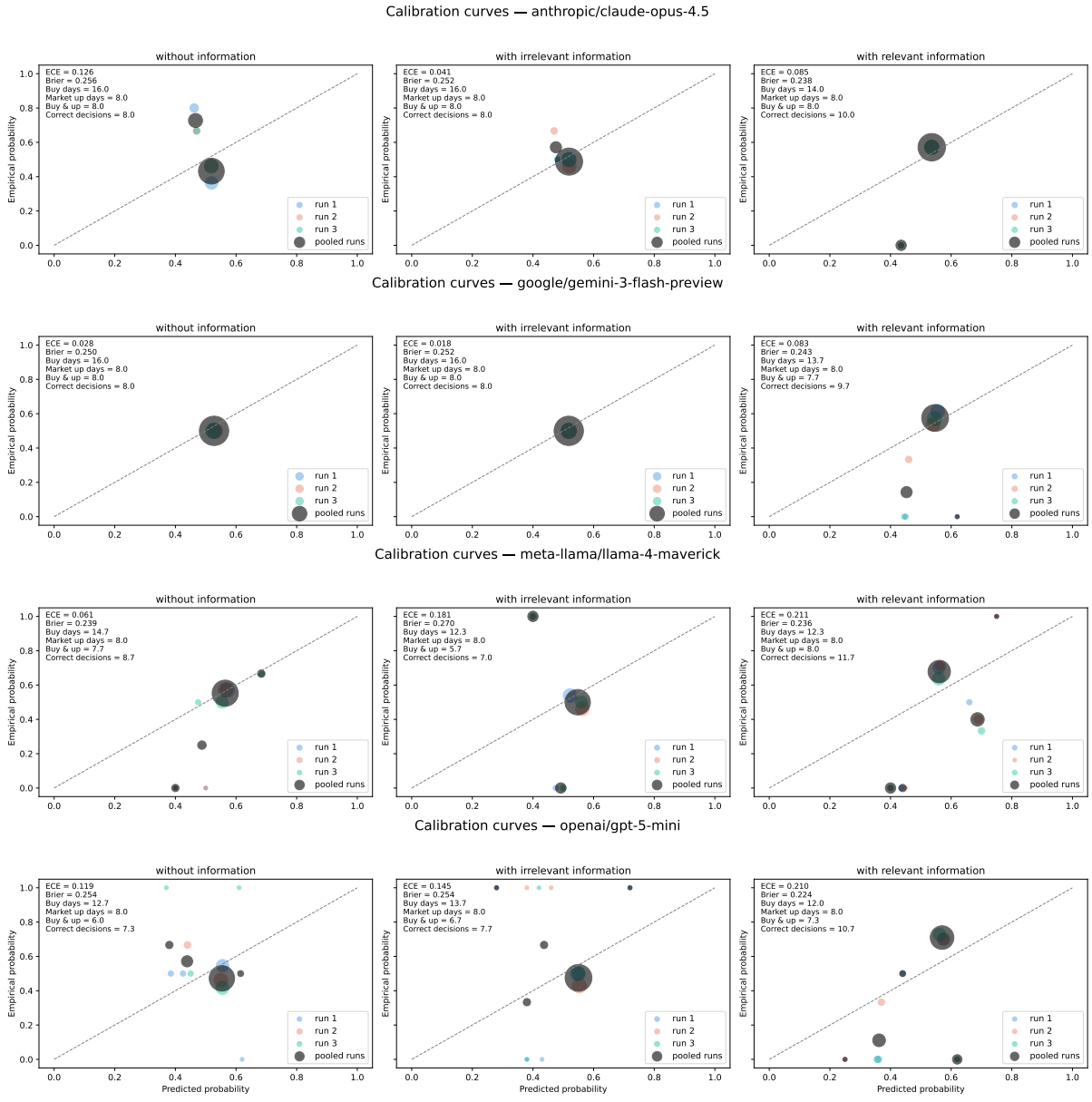

Abbildung 1: Kalibrierungsdiagramme für vier LLMs mit identischen Daten

Trotz der geringen Anzahl an Runs (n = 3) und der damit fehlenden statistischen Signifikanz lassen sich klare qualitative Muster identifizieren, die sowohl mit theoretischen Erwartungen aus der Finanzökonomik als auch mit der psychologischen Overconfidence-Literatur übereinstimmen. Die Analyse erfolgt entlang der drei Treatments (without information, with irrelevant information, with relevant information) sowie entlang der zentralen Bewertungsmaße ECE, Brier Score und diskreter Entscheidungsmetriken.

5.1 Baseline: Without Information

Im Baseline-Treatment, in dem den Modellen ausschließlich der aktuelle DAX-Stand und ihre eigene Entscheidungshistorie zur Verfügung stehen, zeigen alle Modelle eine Konzentration der vorhergesagten Wahrscheinlichkeiten im Bereich um 0.5. Dies ist besonders deutlich in den gepoolten Punkten der Kalibrationsdiagramme sichtbar. Ökonomisch ist dieses Verhalten gut begründbar. In Abwesenheit relevanter Information impliziert die Efficient Market Hypothesis (EMH), dass kurzfristige Kursbewegungen nicht systematisch prognostizierbar sind. Eine konstante Prognose nahe der empirischen Basisrate, also hier etwa 50 % steigende Tage, stellt daher ein rationales Verhalten dar (Fama (1970)). Die beobachteten Brier Scores liegen in diesem Treatment bei allen Modellen in der Nähe von 0.25, was exakt dem Verlust eines uninformierten Prognosemodells bei einer Basisrate von 0.5 entspricht. Gleichzeitig sind die ECE-Werte moderat bis niedrig, insbesondere bei Llama-4-Maverick und Gemini-3-flash, was darauf hindeutet, dass die Modelle ihre Unsicherheit realistisch einschätzen. Diese Ergebnisse sind konsistent mit der Interpretation, dass Overconfidence hier kaum auftritt: Die Modelle zeigen weder Overestimation noch Overprecision, sondern verhalten sich vorsichtig und zurückhaltend. Es liegt keine systematische Fehlkalibrierung, sondern vielmehr ein Zustand rationaler Unsicherheit vor.

5.2 Irrelevante Informationen: Mehr Kontext, keine Verbesserung

Das Treatment mit irrelevanten Informationen führt zu einem deutlich differenzierteren Bild. Während sich die objektive Entscheidungsqualität, gemessen an Correct decisions oder Buy & up, kaum verbessert oder sogar verschlechtert, zeigen mehrere Modelle eine Zunahme der Fehlkalibrierung (Reliability = ECE steigt). Besonders ausgeprägt ist dieser Effekt bei Llama-4-Maverick, wo der ECE von 0.061 im Baseline-Treatment auf 0.181 ansteigt . Gleichzeitig verschlechtert sich der Brier Score, was darauf hindeutet, dass die zusätzliche Konfidenz nicht durch verbesserte Prognoseleistung gedeckt ist. (Resolution$(\rightarrow)$ = Varianz$(\rightarrow)$ + Reliability$(\uparrow)$ − Brier$(\uparrow)$)

Der steigende Brier Score weißt auf Overprecision hin: Die Modelle überschätzen

die Sicherheit ihrer Urteile, obwohl die zugrunde liegende Informationslage objektiv unverändert geblieben ist.

Dieses Ergebnis steht in direkter Übereinstimmung mit der Literatur zu irrelevanter Information und Overconfidence. (Moore and Healy (2008)) argumentieren, dass zusätzliche, aber nicht-diagnostische Information das subjektive Sicherheitsgefühl erhöht, ohne die tatsächliche Genauigkeit zu steigern. Genau dieses Phänomen lässt sich hier beobachten. Die irrelevanten DAX-Fakten wirken wie Informationen, die von den Modellen offenbar als Signal interpretiert werden, obwohl sie keinen prognostischen Gehalt besitzt.

Bemerkenswert ist zudem, dass die Anzahl der Buy-Entscheidungen in diesem Treatment nicht systematisch steigt. Overconfidence äußert sich hier also nicht primär in erhöhter Handelsaktivität, sondern in verschlechterter Kalibration. Dieser Befund, dass die Overconfidence nicht zwingend mit höherem Trading-Volumen gleichzusetzen ist, sondern als Fehlwahrnehmung der eigenen Prognosepräzision interpretiert werden kann, stimmt mit Studien überein (Glaser and Weber (2004)).

5.3 Relevante Informationen: Lernen mit Übertreibung

Mit relevanten Informationen sinkt der Brier Score über alle Modelle hinweg im Vergleich zur Baseline, was auf eine höhere Gesamtgüte der Vorhersagen hinweist. Gleichzeitig steigt jedoch der ECE an, insbesondere bei GPT-5-mini (ECE = 0.210) und Llama-4-Maverick (ECE = 0.211).

Diese Kombination aus sinkendem Brier Score und steigendem ECE weißt aufsteigende Fallunterscheidung der Marktsituaitonen hin. Da die Uncertainty der Marktentwicklung und die Basisrate steigender gain- Tage über Treatments hinweg konstant bleibt, impliziert ein sinkender Brier Score bei gleichzeitig steigender Fehlkalibrierung (Resolution) zwangsläufig eine Zunahme der Resolution.

$(\text{Resolution}(\uparrow) = \text{Varianz}(\rightarrow) + \text{Reliability}(\uparrow) - \text{Brier}(\downarrow))$

Die Modelle sind also besser in der Lage, zwischen unterschiedlichen Marktsituationen zu unterscheiden und ein Richtungssignal zu extrahieren. Dies wird auch durch die steigende Anzahl korrekter Entscheidungen bestätigt, insbesondere im Vergleich zum Baseline-Treatment. Wobei eine korrekte Entscheidung, wie vorher definiert (gain=1 & buy oder gain=0 & sell) bedeutet.

Gleichzeitig zeigen die Kalibrationsdiagramme eine leichte Verschiebung der Punkte nach rechts, ohne einen proportionalen Anstieg der empirischen Wahrscheinlichkeit. Die Modelle sagen also häufiger hohe Wahrscheinlichkeiten voraus, die nicht im gleichen Maße durch tatsächliche Trefferquoten gedeckt sind. Dies ist ein klassischer Ausdruck von Overprecision: Die Modelle lernen ein relevantes Signal, überschätzen jedoch die Präzision dieses Wissens.

Dieses Muster spiegelt zentrale Befunde aus der Overconfidence-Literatur wider. Prims

and Moore (2017) zeigen, dass Overprecision besonders dann auftritt, wenn Akteure glauben, über relevantes Wissen zu verfügen, und ihre Unsicherheit unterschätzen. Auch die Befunde zu Merkle and Schreiber (2025) legen nahe, dass Lernen zwar stattfindet, die Anpassung der Konfidenz jedoch unzureichend bleibt, um vollständige Kalibration zu erreichen. Die vorliegenden Ergebnisse legen nahe, dass Sprachmodelle eine ähnliche Dynamik zeigen: Sie adaptieren ihre Entscheidungen an neue Information, passen ihre Konfidenz jedoch stärker an als ihre tatsächliche Genauigkeit. Overconfidence als Reaktion auf komplexe Aufgaben ist auch kongruent mit der Literatur von Sun et al. (2025). Die vorliegenden Ergebnisse zeigen, dass Overconfidence bei LLMs nicht primär als schlechte Accuracy, sondern als Fehlkalibrierung auftritt.

6 Fazit

Diese Arbeit untersucht, wie Large Language Models probabilistische Entscheidungen im Finanzkontext treffen und wie sich ihre Konfidenz in Abhängigkeit von der Informationsumgebung verändert. Die empirischen Ergebnisse liefern konsistente Evidenz dafür, dass LLMs in ihrer Entscheidungsstruktur der menschlichen ähnelt.

Die Ergebnisse stützen **Hypothese H1**: Im Baseline-Treatment ohne zusätzliche Information verhalten sich die Modelle weitgehend rational im Sinne der EMH. Die vorhergesagten Wahrscheinlichkeiten liegen nahe bei 50 %, die Kalibration ist gut, und die geringe Resolution spiegelt korrekt wider, dass ohne Information keine systematische Prognose möglich ist.

Hypothese H2 wird ebenfalls bestätigt. Irrelevante Informationen verbessern weder die Entscheidungsqualität noch die Trefferquote, führen jedoch in mehreren Modellen zu einer deutlichen Verschlechterung der Kalibration. Dieses Muster ist charakteristisch für Overprecision und steht im Einklang mit der verhaltensökonomischen Literatur, die zeigt, dass zusätzliche, aber nicht-diagnostische Information das subjektive Sicherheitsgefühl erhöht, ohne den Informationsgehalt zu steigern.

Auch **Hypothese H3** findet Unterstützung. Relevante Informationen senken den Brier Score und erhöhen die Anzahl korrekter Entscheidungen, was auf eine höhere Resolution und damit tatsächliches Lernen hindeutet. Gleichzeitig steigt jedoch der Expected Calibration Error deutlich an. Die Modelle überschätzen systematisch die Präzision ihres Wissens und passen ihre Konfidenz stärker an als ihre tatsächliche Prognosegenauigkeit. Lernen findet statt, wird jedoch von Overprecision begleitet.

Insgesamt zeigen die Ergebnisse, dass Overconfidence bei LLMs nicht primär als schlechte Accuracy, sondern als Fehlkalibrierung auftritt. Dies hat wichtige Implikationen für den Einsatz von Sprachmodellen in ökonomischen Entscheidungskontexten: Während LLMs relevante Informationen effektiv verarbeiten können, besteht die Gefahr, dass ihre subjektive Sicherheit überschätzt wird, insbesondere in komplexen und unsicheren Um-

gebungen. Die Kombination aus Brier Score und ECE erweist sich dabei als geeignetes Instrument, um diese Effekte sichtbar zu machen.

A LLM Prompt

```
SYSTEM PROMPT
-------------
You are a decision function.
Output ONLY valid JSON. No explanations, no extra text.
Return exactly one JSON object with this schema:
{"decision":"buy" or "sell","confidence":number between 0.0 and 1.0}


USER PROMPT
-----------
You are a rational investor making a daily decision whether to invest in a
DAX index fund or sell your assets for cash.

Rules:
- You can only choose between "buy" (invest today) or "sell".
- No transaction costs.
- The decision applies for today only.

Information:
- Date: {DATE}
- Today's DAX level: {OPEN_PRICE}
- The prior decisions were {PRIOR_DECISIONS}
- Today's additional information: {EXTRA_INFO}

Task:
- Decide whether to buy or sell.
- Confidence is your subjective probability (between 0.0 and 1.0)
  that tomorrow's DAX close will be higher than today's close.

Return ONLY the JSON object. No extra text!
```

B Zerlegung des Brier Scores

In diesem Appendix wird die formale Zerlegung des Brier Scores in die Komponenten *Uncertainty*, *Resolution* und *Reliability* dargestellt (vgl. Murphy (1973)).

Ausgangspunkt ist die Definition des Brier Scores als erwarteter quadratischer Fehler zwischen der vorhergesagten Wahrscheinlichkeit $\hat{p}$ und dem realisierten binären Ereignis $Y \in \{0, 1\}$:

$$\text{Brier} = \mathbb{E}\big[(\hat{p} - Y)^2\big].$$

Durch Ausmultiplizieren des Quadrats ergibt sich:

$$(\hat{p} - Y)^2 = \hat{p}^2 - 2\hat{p}Y + Y^2.$$

Da für binäre Zufallsvariablen $Y^2 = Y$ gilt, folgt:

$$\text{Brier} = \mathbb{E}[\hat{p}^2] - 2\mathbb{E}[\hat{p}Y] + \mathbb{E}[Y].$$

Nun wird der bedingte Erwartungswert

$$m(\hat{p}) := \mathbb{E}[Y \mid \hat{p}]$$

eingeführt. Mithilfe des Gesetzes der iterierten Erwartung gilt:

$$\mathbb{E}[\hat{p}Y] = \mathbb{E}\big[\hat{p}\,\mathbb{E}[Y \mid \hat{p}]\big] = \mathbb{E}[\hat{p}\,m(\hat{p})].$$

Als nächster Schritt wird $\hat{p}^2$ um $m(\hat{p})$ herum zerlegt:

$$\hat{p}^2 = (\hat{p} - m(\hat{p}))^2 + 2\hat{p}m(\hat{p}) - m(\hat{p})^2.$$

Einsetzen in den Ausdruck für den Brier Score liefert:

$$\text{Brier} = \mathbb{E}\big[(\hat{p} - m(\hat{p}))^2\big] - \mathbb{E}\big[m(\hat{p})^2\big] + \mathbb{E}[Y].$$

Da wiederum

$$\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid \hat{p}]\big] = \mathbb{E}[m(\hat{p})]$$

gilt, kann der verbleibende Term umgeschrieben werden als:

$$\mathbb{E}[Y] - \mathbb{E}\big[m(\hat{p})^2\big] = \mathbb{V}(Y) - \mathbb{V}\big(m(\hat{p})\big),$$

wobei für binäre Y gilt:

$$\mathbb{V}(Y) = \mathbb{E}[Y] - (\mathbb{E}[Y])^2.$$

Damit ergibt sich die bekannte Zerlegung des Brier Scores:

$$\text{Brier} = \underbrace{\mathbb{V}(Y)}_{\text{Uncertainty}} - \underbrace{\mathbb{V}(\mathbb{E}[Y \mid \hat{p}])}_{\text{Resolution}} + \underbrace{\mathbb{E}\big[(\hat{p} - \mathbb{E}[Y \mid \hat{p}])^2\big]}_{\text{Reliability}}.$$

Die Komponente *Uncertainty* beschreibt die irreduzible Unsicherheit des Ereignisses, die allein durch dessen Basisrate bestimmt ist. *Resolution* misst die Fähigkeit des Modells, zwischen Situationen mit unterschiedlichen empirischen Ereigniswahrscheinlichkeiten zu unterscheiden. *Reliability* erfasst systematische Abweichungen zwischen vorhergesagten und tatsächlich realisierten Wahrscheinlichkeiten und entspricht damit Fehlkalibrierung.

Literatur

Barber, B. M. and Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1):261–292.

Binnendyk, J. and Pennycook, G. (2024). Individual differences in overconfidence: A new measurement approach. *Judgment and Decision Making*, 19:1–24.

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

Glaser, M. and Weber, M. (2004). Overconfidence and trading volume. Working paper, University of Mannheim. Working Paper.

Grohmann, A., Menkhoff, L., Merkle, C., and Schmacker, R. (2023). Earn more tomorrow: Overconfidence, income expectations and consumer indebtedness. Discussion Paper 2065, DIW Berlin.

Karki, U., Bhatia, V., and Sharma, D. (2024). A systematic literature review on overconfidence and related biases influencing investment decision making. *Economic and Business Review*, 26(2):130–150.

Klayman, J., Soll, J. B., González-Vallejo, C., and Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79(3):216–247.

Merkle, C. and Schreiber, P. (2025). Learning to be overprecise. *Journal of Business Economics*, 95:467–497.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502–517.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600.

Prims, J. P. and Moore, D. A. (2017). Overconfidence over the lifespan. *Judgment and Decision Making*, 12(1):29–41. Author manuscript, HHS Public Access.

Sun, F., Li, N., Wang, K., and Goette, L. (2025). Large language models are overconfident and amplify human bias. *arXiv preprint*.

Tian, Z., Han, Z., Chen, Y., Xu, H., Yang, X., Xuan, R., Wang, H., and Liao, L. (2026). Overconfidence in LLM-as-a-Judge: Diagnosis and confidence-driven solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence. AAAI 2026.

Xu, C., Wen, B., Han, B., Wolfe, R., Wang, L. L., and Howe, B. (2025). Do language models mirror human confidence? exploring psychological insights to address overconfidence in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25655–25672, Bangkok, Thailand. Association for Computational Linguistics.