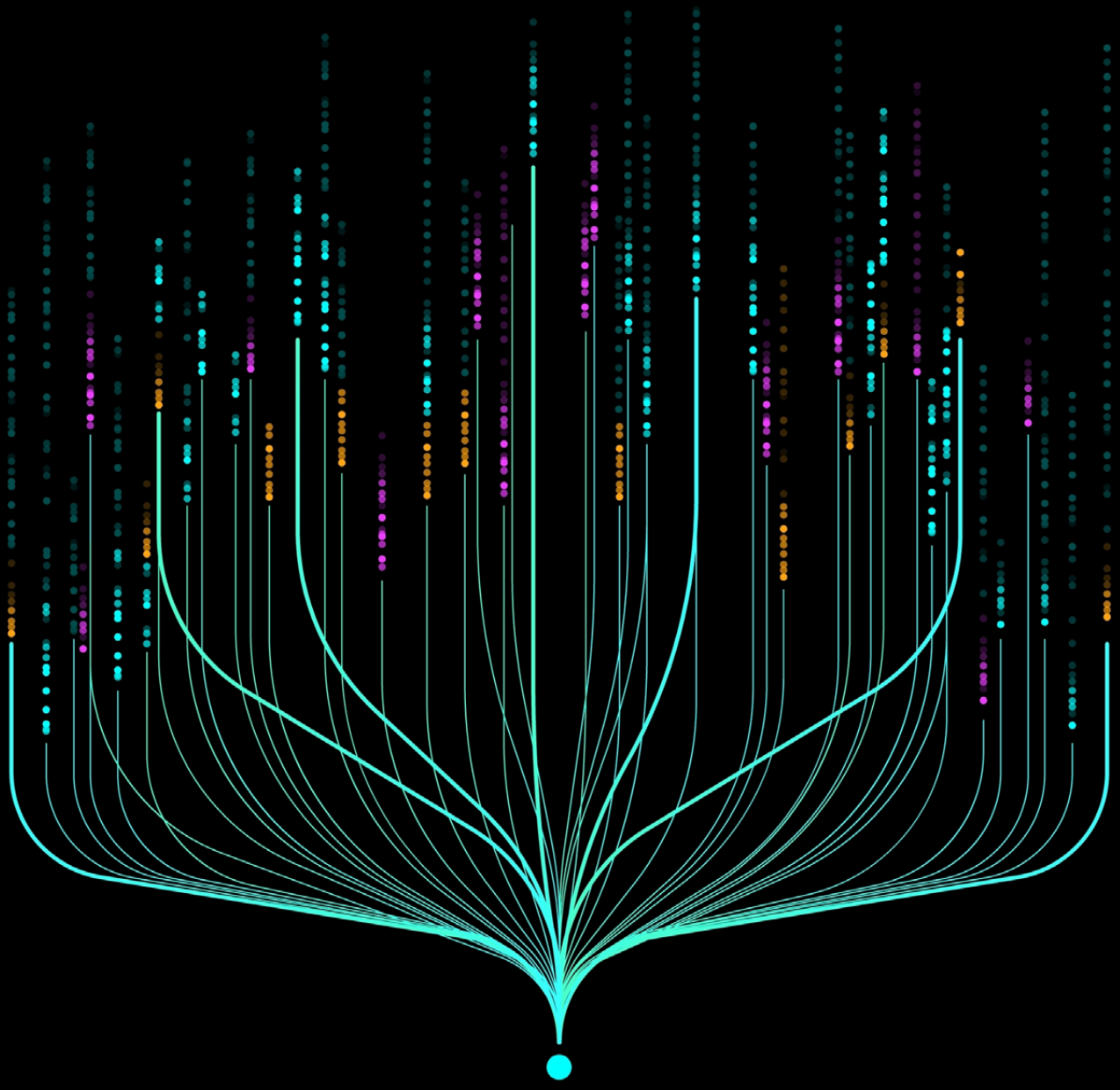

Wie Sprachmodelle bei der Auswertung von Finanztexten helfen können

KI-gestützte Analysen haben großes Potenzial

Die maschinelle Auswertung von Texten mit Hilfe Künstlicher Intelligenz spielt im Asset Management bisher nur eine untergeordnete Rolle. Mit dem Aufkommen großer und leistungsfähiger Sprachmodelle wie ChatGPT könnte sich dies jedoch rasch ändern. Die maschinelle Textanalyse (Natural Language Processing) hat sich in den letzten Jahren rasant entwickelt. Dank der technologischen Fortschritte der letzten Jahre können Algorithmen Texte immer besser verstehen und auswerten. Deka und IQAM Invest forschen daher an Ansätzen zur Generierung textbasierter Investitionssignale.

Um einen Text maschinell verarbeiten zu können, ist es erforderlich, diesen zunächst in eine mathematische Sprache zu „überführen". Während in den frühen 2000er Jahren die Textdarstellung noch auf einfachem „Wörter zählen" (Bag-of-Words-Darstellungen) beruhte, wurde diese zunehmend durch Methoden ersetzt, die auf Wortvektoren, sogenannten „Embeddings", basieren (Mikolov et al. 2013). „Embeddings" sind vieldimensionale Vektorrepräsentationen von Wörtern, die deren semantische Bedeutung repräsentieren. In komplexeren Modellen können Vektoren 768 Dimensionen oder mehr haben. Semantisch ähnliche Wörter sind in bestimmten Bereichen des Vektorraums geballt. So kann aus der geometrischen Lage des Vektors auf eine Ähnlichkeit in der Bedeutung geschlossen werden.

Ein weiterer Meilenstein im Bereich Natural Language Processing war die Veröffentlichung des Transformer-Modells im Jahr 2017 durch Vaswani et al. (2017). Dieses Modell legte den Grundstein für eine Vielzahl moderner NLP-Modelle und ermöglichte eine genauere (mathematische) Darstellung von Texten. Dabei basiert die Idee darauf, dass die Bedeutung eines Wortes durch seinen Kontext definiert wird.

In den darauffolgenden Jahren wurden bedeutende Modelle wie GPT (Generative Pre-trained Transformer) von OpenAI und BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019; Brown et al. 2020) entwickelt. Diese Modelle überzeugen durch ihre Fähigkeit, selbst komplexe Sprachaufgaben zu bewältigen, und setzen neue Maßstäbe beim Textverständnis und bei der Textgenerierung. Seit 2018 hat die Weiterentwicklung großer Sprachmodelle (Large Language Models, LLMs) wie GPT-3 und die Verfeinerung von BERT und deren Nachfolgern die Leistungsfähigkeit von NLP-Systemen weiter gesteigert. LLMs sind in der Regel auf einer riesigen Menge von Texten trainiert, die ihnen ein hochentwickeltes Sprachverständnis und Kontextwissen ermöglichen. Mittlerweile gibt es eine Vielzahl verschiedener LLMs, die sich in Architektur, Parametergröße und Textmenge, die sie verarbeiten können, unterscheiden. Häufige Anwendungsfälle reichen von Textzusammenfassungen über Übersetzungen bis hin zu Frage-Antwort-Systemen (Yang et al. 2024). Neuere wissenschaftliche Untersuchungen kommen sogar zu dem Schluss, dass bestimmte LLMs bereits Merkmale einer allgemeinen Künstlichen Intelligenz aufweisen (Bubeck et al. 2023).

Aktuelle Studie zur Verarbeitung von Earnings Calls mit LLMs

Es ist eine spannende Forschungsfrage, ob moderne Sprachmodelle auch im Portfoliomanagement erfolgreich eingesetzt werden können. Auf den Finanzmärkten werden täglich unzählige Informationen verarbeitet und finden Einfluss in die Preisbildung von Vermögenswerten wie Aktien. Zu den wichtigsten Informationsquellen für Aktienanleger gehören Quartalsberichte und die dazugehörigen Telefonkonferenzen der Geschäftsführung bzw. Vorstandsmitglieder mit Finanzanalysten und Journalisten, den sogenannten „Earnings Calls". In der Studie „What Really Matters in Earnings Calls: Relevance-based Sentiment Analysis with Large Language Models" untersuchen Baur, Wolff und Neumann in einem gemeinsamen Forschungsprojekt von Universität Freiburg und Deka Investment, ob LLMs sinnvolle Handelssignale aus Earnings Calls automatisiert extrahieren können.

Die Autoren verwenden das Open Source-Modell Mixtral der französischen Firma Mistral, ein generatives Sprachmodell mit 47 Milliarden Parametern, da dieses in der Lage ist Textsequenzen von bis zu 32.000 Tokens zu verarbeiten (Jiang et al. 2024) und somit auch längere Earnings Calls vollständig verarbeiten kann.[1] Darüber hinaus schlagen die Autoren eine Filterung der Earnings Calls auf die relevanten Textpassagen vor. Hintergrund ist, dass Earnings Calls in der Regel ein breites Spektrum an Themen beinhalten, von denen einige für den Aktienkurs weniger wichtig oder sogar irrelevant sein können. Diese Passagen könnten die Interpretation durch das LLM erschweren. Die Autoren untersuchen daher, ob eine vorgeschaltete Filterung der Earnings Calls-Transkripte einen Mehrwert bietet.

Daten und Methodik

Als Datenbasis verwenden Baur et al. (2024) 27.846 Transkripte von Earnings Calls für 749 Unternehmen aus dem S&P 500 Index im Zeitraum von 2012 bis 2023. Die Transkripte haben eine durchschnittliche Länge von 3.730 Wörtern. Transkripte mit einer Länge von weniger als 500 Wörtern werden aus dem Datensatz ausgeschlossen. Die zugehörigen Aktienkursdaten für den Zeitraum 2012 bis 2023 werden von Bloomberg bezogen. Den Filtermechanismus zur Extraktion der relevanten Abschnitte aus den Earnings Calls implementieren Baur et al. (2024) wie folgt: Zunächst wird das LLM Mixtral befragt, was generell relevante Informationen in Earnings Calls sind. Die Ausgabesequenz enthält eine Reihe von Schlüsselbegriffen, darunter Finanz- und Ergebniskennzahlen. Diese Ausgabesequenz wird mit dem GTE-Large-Modell (Li et al. 2023), in einen 1.024-dimensionalen Vektor überführt, der den Inhalt des Textes repräsentiert. Als nächstes wird jedes Transkript in Textabschnitte unterteilt. Diese werden ebenfalls mit dem GTE-Large-Modell (Li et al. 2023), in 1.024-dimensionale Vektoren überführt. Anschließend wird für jeden Earnings Call das Ähnlichkeits-Maß „Kosinus-Similarität" zwischen den Vektorrepräsentationen der einzelnen Abschnitte und der Ausgabesequenz mit den relevanten Schlüsselbegriffen berechnet. Nur Textabschnitte mit einer Kosinus-Similarität von 0,7 und mehr werden als relevant eingestuft und in der gefilterten Abfrage verwendet. Die Länge der gefilterten Earnings Calls ist im Durchschnitt 44,5 % kürzer im Vergleich zu den ursprünglichen Transkripten. Um zu überprüfen, ob der Filteransatz tatsächlich funktioniert, nutzen Baur et al. (2024) das Expertenwissen von 45 Portfoliomanagern der Deka Investment. Diese sollen in zufällig ausgewählten Earnings Calls ankreuzen, ob eine Textpassage relevant oder irrelevant für die Aktienkursreaktion ist. Das Ergebnis wird anschließend mit dem Label des automatischen Filteransatzes verglichen. Es zeigt sich eine durchschnittliche Übereinstimmung von über 75 %. Dies zeigt, dass der automatisierte Filteralgorithmus funktioniert und zu ähnlichen Entscheidungen kommt wie ein Portfoliomanager.

Die gefilterten Earnings Calls werden dann an das LLM Mixtral übergeben mit der Anweisung einzuschätzen, ob der (gefilterte) Earnings Call zu einem steigenden („rise") oder fallenden Aktienkurs („fall") führen wird. Um den Auswertungsprozess zu erleichtern, wird das Modell angewiesen nur mit „rise" oder „fall" zu antworten.[2] Die Einschätzung des LLM zur Aktienkursentwicklung wird dann mit der tatsächlichen Kursentwicklung verglichen. Hierzu werden einfache und abnormale (Beta-adjustierte[3]) Aktienrenditen über einen Zeitraum von einem Handelstag berechnet. Um die praktische Umsetzbarkeit der LLM-Signale zu gewährleisen wird für Earnings Calls, die bis zu einer Stunde vor Handelsschluss stattfinden, die abnormale Rendite basierend auf dem darauffolgenden Schlusskurs bis zum Schlusskurs am Folgetag berechnet. Für Earnings Calls, die in der letzten Stunde bis Handelsschluss oder nach Handelsschluss stattfinden, werden die Renditen basierend auf dem Eröffnungskurs am Folgetag (t+1) bis zum Eröffnungskurs am übernächsten Tag (t+2) berechnet.

Ergebnisse

Zunächst analysieren Baur et al. (2024) die Treffergenauigkeit der einfachen und abnormalen Aktienkursrenditen. Zum Vergleich wird die Treffergenauigkeit eines einfachen Sentiment-Ansatzes berechnet, der auf einfachem „Wörter zählen" von positiven und negativen Wörtern beruht. Es werden drei verschiedene Wörterbücher berücksichtigt, die weit verbreitet sind, um das Sentiment zu messen: Die Wörterbücher von Loughran-McDonald (Loughran, McDonald 2011), Henry (Henry 2008) und Harvard IV-4 (Stone 2002). Während Loughran-McDonald und Henry finanzspezifische Wortlisten sind, ist Harvard IV-4 ein allgemeines Wörterbuch. Die Ergebnisse der Klassifikation sind in Tabelle 1 zusammengefasst. Es fällt auf, dass alle Modelle die Kursreaktionen nur mit einer Wahrscheinlichkeit von knapp über 50 % vorhersagen können und damit nur unwesentlich besser abschneiden als einfaches Raten. Dies ist in der hohen Markteffizienz der Aktien im S&P 500 begründet. Bis zum Schlusskurs eine Stunde nach dem Earnings Call bzw. bis zur Markteröffnung am Folgetag hat der Aktienkurs bereits reagiert und die Information aus dem Earnings Call ist zum großen Teil bereits eingepreist. Dennoch fällt auf, dass die Klassifikationsergebnisse des Mixtral-Modells basierend auf gefilterten Earnings Calls mit über 52 % deutlich besser ausfallen als die Ergebnisse der Wörterbuch-Ansätze. Das Filtern auf relevante Textpassagen bringt einen moderaten Mehrwert und verbessert die Klassifikationsergebnisse um rund einen Prozentpunkt. Die Ergebnisse sind für nominale und abnormale Renditen insgesamt sehr ähnlich. Neben der Prognosegüte analysieren Baur et al. (2024) eine Anlagestrategie basierend auf den Prognosen des LLMs. Die Strategie ist einfach. Eine Aktie wird bei einem positiven Signal („rise") gekauft und

1 Verwendet wird das Mixtral 8x7B-Modell in einer 4-Bit-quantisierten Version gemäß dem Generative Post-Training Quantization-Ansatz von Frantar et al. (2023), um die Berechnungen zu beschleunigen. Die Hyperparameter-Optimierung erfolgt mittels Grid Search auf Basis von 400 zufällig ausgewählten Earnings Calls.

2 Der Prompt lautet: "Act as an earnings calls analyst model. Given the following earnings call, please predict the reaction of the stock price on the day following the call. Based on the information provided in the earnings call, answer RISE if the stock price will rise, FALL if the stock price will fall."

3 Beta wird basierend auf täglichen Aktienrenditen und Renditen des S&P 500 Index über die letzten 100 Handelstage vor dem Earnings Call berechnet.

	gefilterte Transkripte (N=27,846)		vollständige Transkripte (N=27,846)	
	Nominal	Abnormal	Nominal	Abnormal
Klassifikationsgüte der Kursreaktion				
Wörterbuch-Ansätze				
Loughran-McDonald	50.34	50.31	50.11	50.22
Henry	50.10	50.08	50.07	50.04
Harvard IV-4	50.01	49.97	50.00	50.00
LLM				
Mixtral	**52.12**	**52.59**	51.33	51.48

Tabelle 1: Klassifikationsgüte (Trefferrate) in %

bei einem negativen Signal („fall") leerverkauft. Die Haltedauer beträgt einen Tag. Analog zur Klassifikationsanalyse wird angenommen, dass die Aktie zum Schlusskurs gekauft wird, wenn der Earnings Call bis zu einer Stunde vor Handelsschluss stattfindet. Findet der Earnings Call bis zu einer Stunde vor Handelsschluss oder nach Handelsschluss statt, so wird angenommen, dass die Aktie zum Eröffnungskurs am Folgetag erworben wird und am Folgetag zur Markteröffnung verkauft wird. Variable Handelskosten werden mit 15 Basispunkten berücksichtigt. Als Benchmark werden eine „Immer Kaufen"-Strategie implementiert, die jede Aktie nach dem Earnings Call kauft und einen Tag hält, sowie eine „Trendfolge"-Strategie, die Aktien basierend auf dem Vortages-Trend kauft oder verkauft. Die Idee dabei ist, dass die Vortagesrendite die initiale Kursreaktion auf die Quartalsergebnisse und den Earnings Call enthält und es sinnvoll sein könnte sich diesem Trend anzuschließen. Die Ergebnisse sind in Tabelle 2 dargestellt:

	gefilterte Transkripte (N=27,846)		vollständige Transkripte (N=27,846)	
	Annualized Return (%)	Sharpe Ratio	Annualized Return (%)	Sharpe Ratio
Long-Short Strategies				
Dictionary Baselines				
Loughran-McDonald	-5.34	-0.18	-5.82	-0.20
Harvard IV-4	-10.04	-0.39	-10.09	-0.39
Henry	-9.43	-0.36	-8.95	-0.34
LLM				
Mixtral	**6.89**	**0.45**	**2.47**	**0.22**
General Baselines				
Immer Kaufen	-9.74	-0.37		
Trendfolge	-14.90	-0.80		

Tabelle 2: Handelsstrategie

Die Ergebnisse der Handelsstrategien zeigen, dass die einfachen Wörterbuchansätze nicht profitabel sind. Die Klassifikationen mit dem LLM Mixtral führen mit den vollständigen Earnings Calls zu einer Rendite von 2,47 % p.a. über den analysierten Zeitraum von 2012 bis 2023. Es zeigt sich ein deutlicher Mehrwert der Filterung der Earnings Calls auf die relevanten Textpassagen. Durch die Filterung steigt die Rendite von 2,47 % p.a. auf 6,89 % p.a., auch die Sharpe Ratio steigt von 0,22 auf 0,44. Die einfachen Benchmark-Strategien „Immer Kaufen" und „Trendfolge" sind ebenfalls nicht profitabel.

Fazit

Die Studie von Baur et al. (2024), die in Kooperation zwischen der Universität Freiburg und der Deka Investment durchgeführt wurde, zeigt, dass LLMs wie das Open Source-Modell Mixtral, in der Lage sind, profitable Handelssignale aus Earnings Calls zu extrahieren. Die Handelssignale weisen eine höhere Prognosegüte auf als einfache Wörterbuch-Ansätze und führen zu einer höheren Performance in einer Handelsstrategie. Zudem führt eine vorgeschaltete Filterung der Earnings Calls auf relevante Textpassagen zu einer verbesserten Klassifikationsgüte und zu einer höheren Rendite im Vergleich zur Nutzung der vollständigen Transkripte. Die Studie zeigt, dass die Nutzung von LLMs im Asset Management einen signifikanten Mehrwert bieten kann, insbesondere durch die gezielte Filterung relevanter Informationen. Zukünftige Forschungen könnten sich darauf konzentrieren, die Modelle weiter zu verfeinern und ihre Anwendung auf andere Arten von Finanzdokumenten und Nachrichtenquellen anzuwenden.

Baur, Katharina, Wolff, Dominik, Neumann, Dirk (2024): What Really Matters in Earnings Calls: Relevance-based Sentiment Analysis with Large Language Models, Working Paper.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. arXiv:2005.14165.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). Gptq: Accurate post-training quantization for generative pre-trained transformers. arXiv:2210.17323.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? Journal of Business Communication, 45 (4), 363–407.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2024). Mixtral of experts. arXiv:2401.04088.Li et al. 2023

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (pp. 3111–3119). La Jolla, CA: Neural Information Processing Systems Foundation.

Stone, P. (2002). General Inquirer Harvard-IV Dictionary. Harvard University, Cambridge, MA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (p. 6000–6010). Red Hook, NY, USA: Curran Associates Inc.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the power of llms in practice: A survey on chatgpt and beyond. ACM Transactions on Knowledge Discovery from Data, 18 (6), 1–32.

KATHARINA BAUR
Universität Freiburg

PROF. DR. DOMINIK WOLFF
Deka Investment GmbH & Frankfurt University of Applied Sciences