Framework zur Bewertung von Open- und Closed-Source-LLMs in der Klassifikation von Nachrichtenmeldungen für das Supply-Chain-Riskmanagement

Axel Wagenitz* , Christina Hofmann-Stölting*, Pia Neumann† , Jan Fischer*
*Hochschule für Angewandte Wissenschaften Hamburg, Department Wirtschaft
†Technische Universität Dortmund, Volkswagen AG

***Zusammenfassung*—Diese Arbeit stellt ein Framework zur systematischen Vergleichsanalyse offener und proprietärer Large Language Models (LLMs) für die Relevanzklassifikation von Nachrichten im Supply-Chain-Risk-Management vor. Es ermöglicht Modellvergleiche ohne annotierte Referenzdaten und liefert quantitative und qualitative Einsichten in Konsistenz, Robustheit und Eignung verschiedener Modelle. In einer exemplarischen Anwendung wurde das Framework genutzt, um 22 LLMs anhand der Relevanzklassifikation von 15.000 Nachrichtenartikeln im Kontext der Automobilindustrie zu analysieren. Die Ergebnisse zeigen, dass leistungsfähige Open-Source-Modelle in Einzelfällen mit kommerziellen Systemen vergleichbar sind und unterstreichen die Bedeutung kontinuierlicher Modellwahl und -evaluation für den praktischen Einsatz.**

***Index Terms*—Large Language Models, Textklassifikation, Evaluationsframework, Interrater-Reliabilität**

I. Einleitung

Unternehmen beobachten Nachrichtenmeldungen regelmäßig, um Hinweise auf mögliche Störungen in ihren Lieferketten frühzeitig zu erkennen. Solche Störungen können zeitverzögert auftreten, etwa durch Naturereignisse, die Produktionskapazitäten beeinträchtigen, oder durch sicherheitsrelevante Vorfälle im Transport, die zu Verzögerungen führen. Die große Zahl potenziell relevanter Meldungen macht eine manuelle Prüfung jedoch sehr aufwändig. Kommerzielle Anbieter bieten zwar Filterlösungen an, diese basieren jedoch auf proprietären Systemen und erlauben in der Regel keine Verarbeitung vertraulicher Unternehmensdaten, da deren Weitergabe mit erheblichen Risiken für IT-Sicherheit und Datenschutz verbunden wäre. Vor diesem Hintergrund untersucht die vorliegende Arbeit, inwieweit Large Language Models (LLMs) für die automatisierte Verarbeitung von Nachrichtenmeldungen geeignet sind. Um Datenabflüsse zu vermeiden, werden ausschließlich offene LLMs berücksichtigt. Das Feintuning solcher Modelle gilt als ressourcenintensiv und erfordert spezialisiertes Expertenwissen, das in vielen Unternehmen nicht verfügbar ist [1]–[4]. Daher setzt die Arbeit auf den Einsatz aktueller vortrainierter Modelle, die ohne Anpassung genutzt und regelmäßig durch neuere Versionen ersetzt werden können, um vom technologischen Fortschritt zu profitieren. Entstanden ist die Untersuchung in Kooperation mit einem Unternehmen der deutschen Automobilindustrie, orientiert sich an realen Praxisanforderungen und bezieht auch Ergebnisse studentischer Arbeiten mit ein. Darauf aufbauend wird im nächsten Abschnitt der Forschungsstand skizziert, um die wissenschaftliche Einbettung der Arbeit zu verdeutlichen.

II. Stand der Forschung

Die Analyse externer Nachrichten für das Supply-Chain-Risk-Management ist seit längerem Gegenstand wissenschaftlicher Untersuchungen. Frühere Arbeiten setzten auf regelbasierte Verfahren oder klassische Information-Retrieval-Methoden, die jedoch nur begrenzt in der Lage waren, komplexe Sprachmuster oder implizite Zusammenhänge zu erfassen. Mit dem Aufkommen neuronaler Sprachmodelle wie BERT verbesserten sich die Möglichkeiten der Ereigniserkennung, oftmals allerdings nur durch aufwändiges Feintuning auf proprietären Datensätzen – mit entsprechend eingeschränkter Reproduzierbarkeit. Neuere Studien belegen zwar, dass LLMs in der Lage sind, risikorelevante Ereignisse in Nachrichten zu erkennen, konzentrieren sich jedoch meist auf den Einsatz einzelner Modelle. Shahsavari et al. [5] entwickeln hierfür ein Framework zur Ereignisextraktion, das stark auf modellindividuelle Ergebnisse zugeschnitten ist. Gilardi et al. [6] zeigen die Leistungsfähigkeit von ChatGPT in Zero-Shot-Szenarien, ohne die Übertragbarkeit auf andere Modelle zu berücksichtigen. Damit bleibt offen, wie stabil und konsistent Ergebnisse über verschiedene Modelltypen hinweg ausfallen – insbesondere im direkten Vergleich von offenen und proprietären Systemen. Genau hier setzt die vorliegende Arbeit an: Sie untersucht den parallelen Einsatz von 22 LLMs, darunter sowohl offene als auch kommerzielle Modelle, und vergleicht deren Ergebnisse systematisch. Der methodische Beitrag liegt in (a) der konsistenzorientierten Evaluation mittels Cohen's Kappa ohne Ground-Truth-Annotation, (b) der Aggregation durch Mehrheitsentscheidungen zur Erhöhung der Robustheit und (c) der Entwicklung einer Infrastruktur, die eine flexible Einbindung beliebig vieler offener Modelle ermöglicht. Besonders relevant ist dabei die Diskussion um offene und proprietäre Systeme. Während offene Modelle durch Feintuning teilweise an die Leistungsfähigkeit kommerzieller Ansätze heranreichen [7], betonen andere Beiträge deren Vorteile hinsichtlich Transparenz, Anpassbarkeit und Datenhoheit [8].

DOI: 10.25968/opus-3789

Vor diesem Hintergrund verfolgt die Arbeit bewusst einen Ansatz ohne Feintuning und mit einem Schwerpunkt auf offenen Modellen. Zum Vergleich werden aber auch proprietäre Modelle berücksichtigt. Aufbauend auf aktuellen Erkenntnissen zu Ensemble-Methoden ([9], [10]) wird dafür eine Infrastruktur entwickelt, die 22 Modelle parallel auswertet und systematisch vergleicht.

III. Forschungsfrage und Zielsetzung

Ausgehend von diesen Überlegungen ergibt sich die zentrale Forschungsfrage: "Kann die Ergebnisqualität bei der Klassifizierung von Nachrichten im Hinblick auf Supply-Chain-Risiken durch den parallelen Einsatz mehrerer offener Large Language Models ohne Feintuning verbessert werden?"

Zur Beantwortung dieser Frage verfolgt die Arbeit vier Teilziele:

1) den Vergleich offener LLMs bei identischem Prompt,
2) die Analyse von Mehrheitsentscheidungen als Methode zur robusteren Klassifikation,
3) die Konzeption und Darstellung einer Infrastruktur zur parallelen Modellnutzung,
4) einen Ausblick auf die Integration interner Unternehmensdaten für eine präzisere Einordnung.

IV. Methodik

Zur Beantwortung der Forschungsfrage wurde ein mehrstufiges Vorgehen gewählt, das sowohl die technische Umsetzung der Multi-LLM-Infrastruktur als auch deren systematische Evaluation umfasst.

A. Datenbasis und Klassifikationsaufgabe

Die Datenpipeline bildet die erste Verarbeitungsstufe und gewährleistet über eine einheitliche Schnittstelle das konsistente Einlesen unterschiedlichster Textquellen aus Nachrichtenfeeds, internen Reports oder externen Datenbanken. Zunächst erfolgt eine umfassende Normalisierung, bei der nicht druckbare Zeichen entfernt, verschiedene Zeichensätze auf Unicode vereinheitlicht und Stopwörter herausgefiltert werden. Anschließend werden die Texte in ein standardisiertes JSON-Schema überführt, das Metadaten, Rohtext und Vorverarbeitungsstatus enthält. Durch die strikte Trennung dieser Vorverarbeitung von der Modellanfrage lassen sich neue Datenquellen oder zusätzliche Reinigungsstufen problemlos integrieren, ohne dass Anpassungen am Kernsystem erforderlich werden.

B. LLM-Orchestrierung

Die Auswertung erfolgte mit 22 aktuellen offenen und proprietären Large Language Models, die über ein zentrales Orchestrierungssystem angesprochen wurden. Identische Zero-Shot-Prompts stellten sicher, dass alle Modelle unter gleichen Bedingungen getestet wurden. Eine modulare Adapterstruktur erlaubte die flexible Anbindung unterschiedlicher Modellanbieter. Zur Sicherstellung der Validität integrierte die Infrastruktur Mechanismen zur Fehlererkennung (z. B. Parsing-Fehler, leere Antworten) und Wiederholung fehlerhafter Anfragen.

C. Evaluationsansatz

Eine manuelle Annotation des gesamten Korpus von 15.000 Artikeln wäre aufgrund des erheblichen Aufwands nicht realisierbar und ist für den vorliegenden Ansatz methodisch auch nicht erforderlich, da kein Feintuning der Modelle vorgenommen wird. Die Modellgüte wird daher nicht anhand klassischer Gütemaße wie Accuracy oder F1-Score beurteilt, sondern über die Konsistenz zwischen den Modellen selbst.

1) Evaluationsmaß: Im Zentrum steht Cohen's Kappa als etabliertes Maß der Interrater-Reliabilität. Es quantifiziert die Übereinstimmung zweier "Rater"– hier zweier LLMs – und korrigiert diese um die Wahrscheinlichkeit zufälliger Übereinstimmungen [11]. Damit ist Kappa besonders geeignet, da keine annotierte Ground-Truth vorliegt und reine Übereinstimmungsraten die Konsistenz überschätzen würden. Die Berechnung erfolgt paarweise über sämtliche Modellkombinationen hinweg. So lässt sich aufzeigen, in welchem Ausmaß bestimmte Modelle systematisch übereinstimmen oder voneinander abweichen. Hohe Kappa-Werte weisen auf konsistente Klassifikationslogiken hin, niedrige Werte auf divergente Bewertungsmuster.

2) Aggregationsstrategien: Während Cohen's Kappa die Konsistenz zwischen Modellen misst und damit ein Evaluationsmaß darstellt, sind Aggregationsstrategien darauf ausgerichtet, aus den unterschiedlichen Klassifikationen ein praktisches Gesamtergebnis abzuleiten. Als erster Ansatz wird die Mehrheitsmeinung genutzt: Für jede Nachricht wird das am häufigsten vergebene Relevanzlabel über alle Modelle bestimmt. Dieses Verfahren stellt eine baseline dar, die in weiteren Arbeiten durch gewichtete Mehrheiten und Unsicherheitsmetriken erweitert werden soll.

D. Architektur

Die entwickelte Infrastruktur folgt einem modularen Aufbau mit klar getrennten Komponenten für Datenpipeline, Modellorchestrierung und Evaluation. Alle Zwischenschritte, einschließlich Modellantworten, Protokollen und Fehlermeldungen, wurden in einer relationalen Datenbank gespeichert, sodass vollständige Nachvollziehbarkeit gewährleistet ist. Durch dieses methodische Vorgehen können die Ergebnisse sowohl im Hinblick auf technische Reproduzierbarkeit als auch auf wissenschaftliche Validität eingeordnet werden.

V. Ergebnisse

Die Auswertung der 22 LLMs zeigt deutliche Unterschiede in der Klassifikation. Während einige Modelle konservativ bewerten und den Großteil der Nachrichten als nicht relevant einordnen, neigen andere zu einer höheren Sensitivität. Bei rund 92,15 % der Nachrichtenmeldungen lag eine Übereinstimmung zwischen mindestens 15 der 22 Modelle vor. In 3,71 % ergaben sich klare Mehrheitsentscheidungen (50 % der abgegebenen Ratings haben dasselbe Label), während 4,13 % eine hohe Varianz (keine klare Mehrheit) zeigten. Die Mehrheitsaggregation führte zu stabileren Ergebnissen und

Tabelle 1. Ueberblick ausgewählter LLMs

Modellname	Veröffentlichung	Lizenz	MoE	Modellfamilie	Parameter	Wissensstand
kimi-k2	11. Jul 2025	Offen	Ja	Kimi K2	1 T, 32 B aktiv	Juni 2025[a]
claude-3-5-haiku-20241022	22. Okt 2024	Proprietär	Unbekannt	Claude 3.5	Unbekannt	Juli 2024
claude-sonnet-4-20250514	22. Mai 2025	Proprietär	Unbekannt	Claude 4	Unbekannt	Mrz 2025
deepseek-v3	20. Jan 2025	Offen	Ja	DeepSeek-V3	671 B, 37 B aktiv	Juli 2024[a]
deepseek-r1	28. Mai 2025	Offen	Ja	DeepSeek-R1	671 B, 37 B aktiv	Oktober 2023[a]
gemini-2.0-flash-001	05. Feb 2025	Proprietär	Unbekannt	Gemini 2.0	Unbekannt	Juni 2024
gemini-2.5-flash-001	17. Jun 2025	Proprietär	Unbekannt	Gemini 2.5	Unbekannt	Jan 2025
gemma-3-12b	11. Mrz 2025	Offen	Nein	Gemma 3	12 B	August 2024
gemma-3-27b	11. Mrz 2025	Offen	Nein	Gemma 3	27 B	August 2024
gpt-4.1-mini-2025-04-14	14. Apr 2025	Proprietär	Unbekannt	GPT-4.1	Unbekannt	Juni 2024
gpt-4o	20. Nov 2024	Proprietär	Unbekannt	GPT-4o	Unbekannt	Juni 2024
gpt-4o-mini	18. Jul 2024	Proprietär	Unbekannt	GPT-4o	Unbekannt	Oktober 2023
llama-4-maverick	05. Apr 2025	Offen	Ja	LLaMA 4	Unbekannt	August 2024
llama3.3	06. Dez 2024	Offen	Nein	LLaMA 3	70 B	Dezember 2023
mistral-nemo	16. Jul 2024	Offen	Nein	Mistral Nemo	12 B	April 2024
mistral-small-3.1-24b	17. Mrz 2025	Offen	Nein	Mistral Small 3.1	24 B	Dezember 2024[a]
phi-4	12. Dez 2024	Offen	Nein	Phi 4	14 B	Juni 2024
qwen3-14b	29. Apr 2025	Offen	Nein	Qwen3	14 B	Oktober 2023
qwen3-235b-a22b	29. Apr 2025	Offen	Ja	Qwen3 MoE	235 B, 22 B aktiv	Oktober 2023
qwen3-30b	29. Apr 2025	Offen	Ja	Qwen3 MoE	30 B, 3 B aktiv	Oktober 2023
qwen3-32b	29. Apr 2025	Offen	Nein	Qwen3	32 B	Oktober 2023
qwen3-8b	29. Apr 2025	Offen	Nein	Qwen3	8 B	Oktober 2023

[a] vermutet

verringerte extreme Fehlklassifikationen. Über alle Meldungen ergab die Mehrheitsaggregation folgende Verteilung:

- nicht relevant: 96,05 %
- Relevanz niedrig: 0,12 %
- Relevanz mittel: 0,11 %
- Relevanz hoch: 0,16 %

Die Varianz der Antworten erwies sich zudem als Indikator für Unsicherheit und kann Unternehmen auf Nachrichten hinweisen, die einer genaueren Betrachtung bedürfen. Anhand Abbildung 1 lassen sich bei der Relevanzbewertung Bewertungscluster und Ausreißer erkennen. Mit einer maximalen Inter-Rater Konsistenz von 58 % kommen Modelle innerhalb einer Modellfamilie am stärksten zu einer identischen Relevanzeinschätzung. Dieser Umstand ist aufgrund ähnlicher Architekturen und Trainingsdaten innerhalb von Modellfamilien nicht unerwartet. Anhand Cohen‘s Kappa zeigt sich jedoch weiterhin, dass auch kostenpflichtige und kostenfreie LLMs mit bis zu 57 % eine vergleichbare Inter-Rater-Konsistenz aufweisen können. Je nach Wahl der Modelle gibt es somit keinen signifikanten Unterschied zwischen proprietären und open-source Varianten. So erreichen die Modelle von Meta (llama3.3 und llama-4-maverick) nahezu identische Kappa Werte wie die OpenAI LLMs mit seinem State-of-the-art-Modell gpt-4o. Die ähnlichen Bewertungsmuster zwischen proprietären und offenen Modellen deuten auf eine funktionelle Angleichungen der LLMs, beispielsweise durch ähnliche Transformer-Architekturen oder Trainingsmethoden, hin. Das Ergebnis ist aus unternehmerischer Sicht ebenso interessant, da die Ergebnisgüte eines entsprechenden Systems somit nicht von den Kosten der verwendeten Modelle abhängt. Stattdessen kann auch mit einem Ensemble kostenloser Modelle ein vergleichbares Ergebnis erreicht werden. Mit Hilfe von Cohen‘s Kappa lassen sich darüber hinaus Ausreißer identifizieren, welche die durchschnittliche Übereinstimmungsrate erheblich beeinflussen. Die Gemma Modelle und gemini-2.0-flash kommen in dem betrachteten Anwendungsfall systematisch zu abweichenden Risikobewertungen. Am größten ist der Dissens dabei zwischen gemma-3-12b und qwen3-30b mit nur 8% Übereinstimmung in der Risikobewertung. Diese hohen Abweichungen unterstreichen erneut die Abhängigkeit des Verhaltens großer Sprachmodelle von entwicklungsbedingten Faktoren wie den Trainingsdaten oder der Tokenisierung, was dazu führt, dass Modelle konservativ oder optimistisch bewerten [12], [13].

Mit einem durchschnittlichen Kappa von 0,361 bei einer Standardabweichung von 0,182 unterstreichen die Ergebnisse insgesamt, dass in dem betrachteten Anwendungsfall unter den 20 Modellen keine homogene Mehrheitsmeinung existiert. Während gewisse Modellgruppen sehr ähnliche Bewertungen liefern, haben andere Modellgruppen wiederum konträre Meinungen hinsichtlich der Relevanz von bestimmten Nachrichten für die Versorgungssicherheit eines Unternehmens. Ein sorgfältig ausgewähltes Ensemble aus Modellen verschiedener Anbieter hat jedoch das große Potenzial, die Robustheit der Relevanzbewertung insgesamt zu erhöhen. Schließt man nur die drei größten Ausreißer aus dem Ensemble aus, sinkt die Standardabweichung bereits auf 0,093 und der Cohen‘s Kappa Durchschnitt steigt auf 0,39.

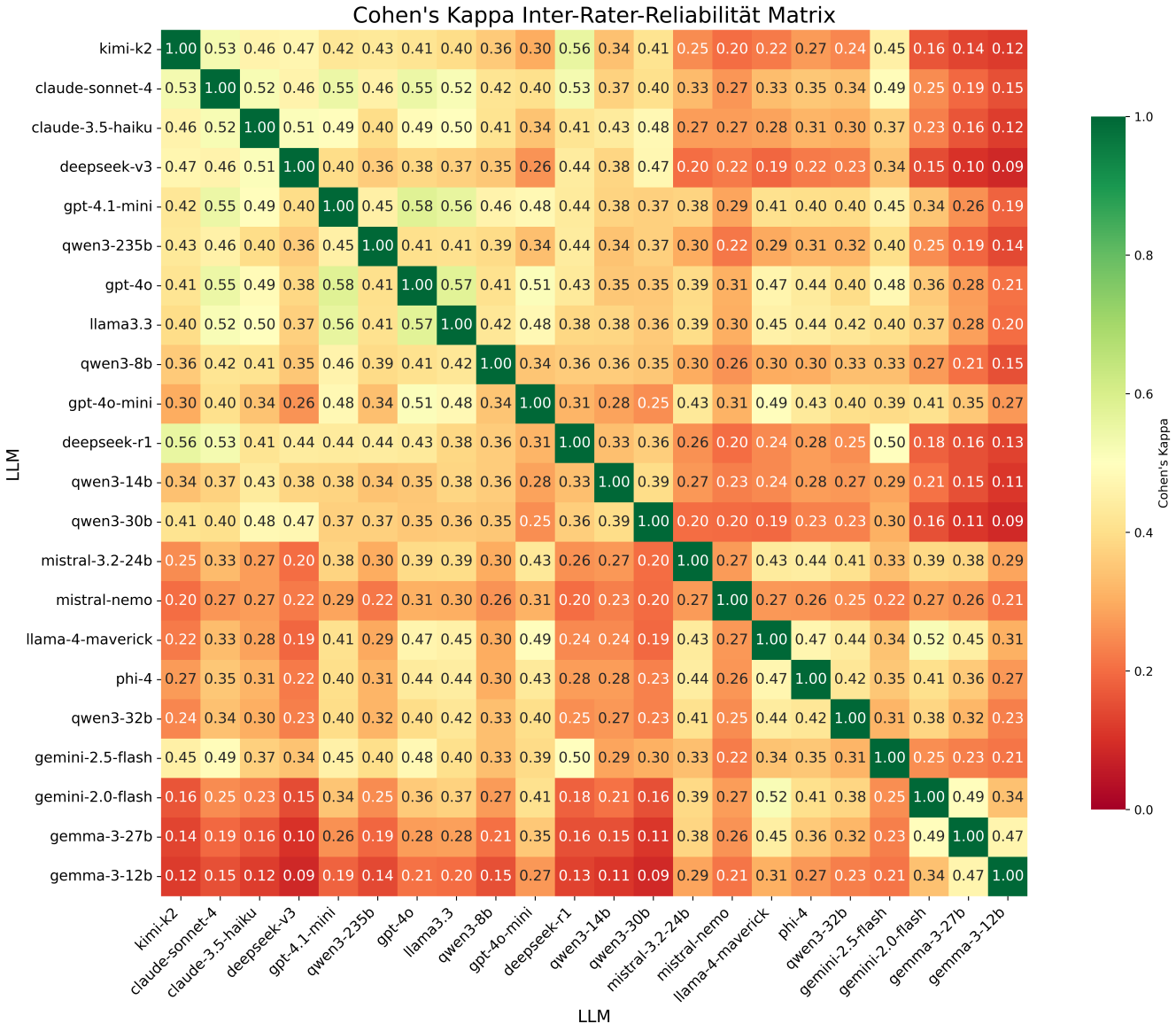

Abbildung 1. Cohen‘s Kappa Heatmap

VI. Diskussion und Ausblick

Die Ergebnisse zeigen, dass der parallele Einsatz mehrerer LLMs eine robuste Alternative zur Einzelmodellbewertung darstellt. Mehrheitsentscheidungen reduzieren Ausreißer, Varianzanalysen liefern Hinweise auf Unsicherheiten. Die entwickelte Infrastruktur gewährleistet zudem Nachvollziehbarkeit, da sämtliche Modellantworten gespeichert und retrospektiv analysierbar sind. Damit entsteht eine flexible Grundlage, die technisch erweiterbar und wissenschaftlich reproduzierbar ist.

Für den praktischen Einsatz bestehen jedoch Grenzen: Die Ergebnisse beruhen auf allgemeinen Relevanzbewertungen und erlauben noch keine kontextspezifischen Handlungsempfehlungen. Erst die Verknüpfung mit internen Unternehmensdaten, die Definition klarer Eskalationslogiken und die Integration in bestehende IT-Systeme kann zu einem operativ nutzbaren Instrument im Supply-Chain-Risk-Management führen. Darüber hinaus bietet der Ansatz Potenzial für die Lehre. Unterschiedliche Textgrundlagen, wie technische Dokumente oder juristische Texte, können eingesetzt werden, um Studierenden Methoden wie Konsistenzanalyse und Ensemble-Verfahren praxisnah zu vermitteln. Die nächsten Schritte umfassen die Anbindung unternehmensinterner Datenquellen zur kontextsensitiven Bewertung, die Weiterentwicklung geeigneter Aggregationsstrategien sowie die Erprobung und Validierung der Ansätze in praktischen wie auch didaktischen Anwendungsszenarien.

Transparenzhinweis

In diesem Text wurden Large Language Modelle als unterstützende Werkzeuge zur sprachlichen Ausformulierung und

Strukturierung verwendet. Inhaltliche Aussagen, Zielsetzungen und methodische Entscheidungen liegen in der Verantwortung der Autor:innen.

ANHANG

Auflistung 1. System Prompt

```
system_prompt = ("You are an expert in
    analyzing risks of disruptions in the
    automotive "
"supply chain. Your primary task is to assess
    whether a given news article "
"indicates a potential or actual interruption
    in the supply of automotive "
"parts. Consider geopolitical events, natural
    disasters, labor strikes, "
"factory shutdowns, logistics issues, and any
    other relevant factors. "
"Base your analysis on factual cues and
    logical inference.")
```

Auflistung 2. Base Prompt

```
base_prompt = ( "Analyze the following news
    article to determine whether it has a
    severe "
"impact on the automotive supply chain.
    Clearly structure your reasoning by "
"enclosing your entire reasoning process
    between explicit tags: <think> and "
"</think>.\n\n"
"After the reasoning section, output ONLY a
    JSON snippet with the following "
"structure:\n"
"{\n"
        '  \"relevance\": \"RELEVANCE_NO\" |
    \"RELEVANCE_LOW\" | \"RELEVANCE_MEDIUM\"
    | \"RELEVANCE_HIGH\",\n'
        '  \"disruption_timeline\": \"SHORT\"
    | \"MEDIUM\" | \"LONG\"\n'
        "}\n\n"
"Definitions:\n"
"- 'relevance': indicates the severity of the
    impact on the automotive supply chain.\n"
"- 'disruption_timeline': indicates when the
    disruption is expected to start.\n"
"    - SHORT: within six weeks\n"
"    - MEDIUM: within twelve weeks\n"
"    - LONG: after twelve weeks or
    longer-term effects\n\n"
"IMPORTANT:\n"
"- If there is no direct impairment, set
    relevance to 'RELEVANCE_NO' and leave
    disruption_timeline empty.\n"
"- Do NOT include anything outside of the
    <think>...</think> reasoning and the
    final JSON snippet.\n"
"- NEVER leave the JSON response empty.\n")
```

LITERATUR

[1] LearningDaily, „What is the cost of fine-tuning LLMs?," 2025. URL: https://learningdaily.dev/what-is-the-cost-of-fine-tuning-llms-f5801c00b06d.

[2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, „LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2023. DOI: 10.48550/arXiv.2106.09685.

[3] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, „Parameter-Efficient Transfer Learning for NLP," in *International Conference on Machine Learning (ICML)*, 2019, pp. 2790–2799. DOI: 10.48550/arXiv.1902.00751.

[4] N. Ding, Y. Chen, X. Zhang, Z. Chen, Y. Qin, Z. Zhang, Z. Liu, M. Sun, and J. Tang, „Delta Tuning: A Comprehensive Study of Parameter Efficient Methods for Pre-trained Language Models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1198–1212, 2022. DOI: 10.48550/arXiv.2203.06904.

[5] M. Shahsavari, O. K. Hussain, M. Saberi, and P. Sharma, „Event Identification for Supply Chain Risk Management Through News Analysis by Using Large Language Models," *The Review of Socionetwork Strategies*, vol. 18, pp. 255–278, 2024. DOI: 10.1007/s12626-024-00169-z

[6] F. Gilardi, M. Alizadeh, and M. Kubli, „ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks," *PNAS*, vol. 120, no. 30, art. e2305016120, 2023. DOI: 10.1073/pnas.2305016120.

[7] G. Zhang et al., „Closing the Gap Between Open Source and Commercial Large Language Models for Medical Evidence Summarization," *NPJ Digital Medicine*, vol. 7, art. 239, 2024. DOI: 10.1038/s41746-024-01239-w.

[8] J. Manchanda et al., „The Open-Source Advantage in Large Language Models (LLMs)," *arXiv preprint*, 2024. DOI: 10.48550/arXiv.2412.12004.

[9] Z. Chen et al., „Harnessing Multiple Large Language Models: A Survey on LLM Ensemble," *arXiv preprint*, 2025. DOI: 10.48550/arXiv.2502.18036.

[10] Y. Huang et al., „Ensemble Learning for Heterogeneous Large Language Models with Deep Parallel Collaboration," in *Advances in Neural Information Processing Systems*, vol. 37, 2024. DOI: 10.48550/arXiv.2404.12715.

[11] J. Cohen, „A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. DOI: 10.1177/001316446002000104.

[12] M. A. Desai, I. V. Pasquetto, A. Z. Jacobs, and D. Card, „An archival perspective on pretraining data," *Patterns*, vol. 5, no. 4, art. 100966, 2024. DOI: 10.1016/j.patter.2024.100966.

[13] R. Sennrich, B. Haddow, and A. Birch, „Neural Machine Translation of Rare Words with Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162.