Sequencing News Articles with Large Language Models within Enterprise Risk Management Context

Žiga Debeljak†
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
ziga.debeljak@mps.si

Dunja Mladenić
Department for Artificial Intelligence,
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

Klemen Kenda
Department for Artificial Intelligence,
Jožef Stefan Institute
Ljubljana, Slovenia
klemen.kenda@ijs.si

Abstract

This paper evaluates the capability of Large Language Models (LLMs) to reconstruct event timelines from unstructured news data. This capability is highly relevant for Enterprise Risk Management (ERM) applications, where the reconstruction and forecasting of coherent event trajectories are crucial for identifying, assessing, and predicting emerging risks and analyzing risk scenarios. In this study, we tasked twenty LLMs with chronologically ordering randomly shuffled business news articles for three distinct real-world event chains. To prevent simple date sorting, all explicit date markers were removed from the articles. The experiments were conducted under one unassisted and three assisted scenarios that provided the models with hints for the first, the last, or both the first and the last articles in the sequence. The results reveal a systematic variation in difficulty across the three tasks in addition to significant performance disparities among the models, with Grok 4 (xAI), GPT-5, o3 and o3-pro (all three OpenAI), and Gemini 2.5 Pro (Google) consistently outperforming other models practically across all tasks and prompting scenarios. As expected, prompting assistance with additional information systematically improved accuracy, especially for the models that performed poorly in the unassisted scenario. The high level of accuracy achieved by the top-performing models indicates a practical utility for real-world ERM applications.

Keywords

Large Language Models, News-Stream Sequencing, Temporal Reasoning

1 INTRODUCTION

Within Enterprise Risk Management (ERM) practice, organizations monitor external developments also by analyzing streams of publicly available news. Each news article captures a momentary state of the political-economic environment, and by accurately structuring unordered information into a chronological narrative, organizations can better understand the evolution of events and the relationships that connect them. The reconstruction and forecasting of these event trajectories are important for identifying, assessing, and predicting emerging risks, especially within risk scenario analysis [10, 11]. The capability to build structured timelines from unstructured textual information is therefore of high relevance to ERM.

LLMs are increasingly utilized in ERM for their ability to process and analyze unstructured textual data, including news articles, to identify and assess risks [1, 2, 3, 4, 5]. In the financial sector, applications include extracting sentiment from news to gauge market perception or identify reputational risks [3, 6, 7, 8], and identifying specific risk factors or events discussed in news and corporate disclosures [2, 4, 5, 9]. Existing literature mainly demonstrates LLMs' utility in analyzing individual or aggregated news items for tasks such as sentiment analysis, risk factor identification, or event detection, but the capabilities of the models to recover the temporal order and causal links among a sequence of discrete news items that describe an unfolding narrative are less directly explored. This paper aims to address this gap by investigating LLM performance in temporal-causal reasoning within news streams, a crucial aspect for understanding the dynamics of unfolding risk narratives.

By investigating whether state-of-the-art commercial or open-source LLMs can reconstruct the chronological narrative of business-event chains from unordered news articles, this paper contributes to the field by: (a) systematically evaluating the performance of multiple LLMs on a challenging temporal-reasoning task; (b) analysing the efficacy of diverse prompting strategies — both unassisted and assisted — in improving model accuracy; (c) providing insights into model-and-task dynamics, revealing substantial performance disparities, task-specific difficulty patterns, and the outsized gains weaker models receive from contextual hints; and (d) demonstrating the practical readiness of these technologies for ERM deployment.

2 RESEARCH METHOD

Task Definition

To evaluate the capabilities of LLMs, three event chains were constructed, focusing on: (1) Trump's Tariffs and EU [“Task_1”], (2) Gold Prices [“Task_2”], and (3) the Ukraine-Russia War [“Task_3”]. These topics were selected due to their significant relevance to the business environment. For each topic, ten articles were manually selected from the online editions of two reputable sources of financial and business information, published between March 1st and May 2nd, 2025. For the purpose of LLM processing, the raw text from the selected articles was extracted. To prevent temporal bias, explicit date indicators—such as full dates—were removed, and no two

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

http://doi.org/ 10.70314/is.2025.sikdd.4

articles shared the same publication date. Subsequently, the articles within each event chain were randomly shuffled, and this fixed random order was then applied to all models within the experiment.

The primary task for the selected LLMs was to reconstruct the chronological sequence of news articles within three distinct event chains. This task was evaluated across four experimental scenarios: (1) an unassisted scenario ["Assist_No"], and three assisted scenarios providing the (2) first ["Assist_First"], (3) last ["Assist_Last"], or (4) both first and last ["Assist_FirstLast"] articles in the sequence.

In the unassisted scenario, the LLMs were required to determine the correct chronological order of the articles without any external information regarding their placement. In the assisted scenarios, the models were provided with hints within the user prompt. Specifically, for the Assist_First and Assist_Last scenarios, the prompt identified the article occupying the initial or final position, respectively. In the Assist_FirstLast scenario, the LLMs were given the identifiers for the articles that correspond to the beginning and end of the chronological sequence.

The required output from the LLMs was a reconstructed timeline of the news articles. For each position in the timeline, the following information was mandated: (i) the article's identification number, (ii) the article's title, (iii) a brief justification for its placement relative to the preceding article, and (iv) a brief justification for its placement relative to the subsequent article. The models were required to provide a structured output in JSON format.

Prompt Engineering

Prompt engineering included manual drafting, testing on different models, and optimization both with LLMs (GPT o3 and Gemini 2.5 Pro) as well as manually, in several iterations. In the end, an effective user prompt was developed which worked reasonably well for all selected models. The main challenges with regard to the design of prompts were: (a) stimulating a systematic approach to causal reasoning, which was considered to be mainly important for the non-reasoning models; (b) ensuring the output consisted of exactly ten distinct articles, with no repetitions or omissions; (c) enforcing the required output JSON schema; and (d) providing concise reasoning for the positioning of the observed articles.

Within the user prompt, the models were explicitly instructed to use the following reasoning principles: (a) inferring sequences of events (how events described in different articles relate to each other over time), (b) causal reasoning (identifying cause-and-effect relationships between the content of different articles), (c) logical story progression (understanding how a narrative or situation typically develops or unfolds), (d) utilizing any implicit time references if available within the articles, and (e) using models' general knowledge about events. Prompts with clear instructions about the guidelines for the reasoning process worked better than prompts without such instructions, even with models with strong reasoning capabilities. System prompts were not utilized, as the one-shot user prompt contained all necessary instructions for the models. The full user prompt is available from the authors.

Selected LLMs and Experiment Execution

Twenty different models by eight different providers were selected for this research, based on their expected capabilities with regard to the tasks, and their availability. Overview of selected models is shown in Table 1.

Table 1: Selected LLMs

#	Model Provider: Model Name	Context Window (tokens)	Date Created
1	OpenAI: GPT-4.1	1.047k	14.04.2025
2	OpenAI: o3	200k	16.04.2025
3	OpenAI: o3-pro	200k	10.06.2025
4	OpenAI: gpt-oss-120b	131k	5.08.2025
5	OpenAI: GPT-5	400k	7.08.2025
6	Google: Gemini 2.5 Pro Preview	1.048k	7.05.2025
7	Google: Gemini 2.5 Flash Preview	1.048k	20.05.2025
8	xAI: Grok 3 Beta	131k	9.04.2025
9	xAI: Grok 4	256k	9.07.2025
10	Anthropic: Claude Sonnet 4	200k	22.05.2025
11	Anthropic: Claude Opus 4	200k	22.05.2025
12	Anthropic: Claude Opus 4.1	200k	5.08.2025
13	Meta: Llama 4 Maverick	1.048k	5.04.2025
14	Meta: Llama 4 Scout	1.048k	5.04.2025
15	Mistral AI: Mistral Medium 3	131k	7.05.2025
16	Mistral AI: Mistral Medium 3.1	262k	13.08.2025
17	Qwen: QwQ 32B	131k	5.03.2025
18	Qwen: Qwen 2.5 VL 32B Instruct	128k	24.03.2025
19	DeepSeek: DeepSeek V3	163k	24.03.2025
20	DeepSeek: R1	128k	28.05.2025

All models were accessed using the OpenRouter platform via the APIs. For models supporting this parameter, the temperature was set to 0.0 to ensure the most reliable and reproducible experimental results; otherwise, default parameters were used.

There were 12 experiments executed: 3 different event topic chains (tasks) in 4 experimental scenarios (prompts) each, by using all 20 LLMs as shown in Table 1, thus resulting in 240 results (outputs). Experiments were executed on June 1st, 2025 with the models available on that date, and on August 19th, 2025 with the newer models.

3 EVALUATION AND DISCUSSION

General Evaluation

In terms of the **output content**, all models demonstrated strong performance in response to a standardized user prompt, successfully producing the requested ordered lists of news articles with all accompanying metadata. From a logical standpoint, the outputs from all models were accurate, presenting ordered lists that included all required supplementary information. Substantial variations in output quality were observed across the different models. This variation was also influenced by the three distinct tasks, which seemed to be of substantially different difficulty, with the first task being the most straightforward and the last presenting the most significant challenge. As anticipated, the implementation of assisted prompting strategies consistently enhanced the accuracy of the outputs for all models across all evaluated tasks.

Regarding the **output formatting**, the majority of the models adhered to the specified JSON schema. Notable exceptions to

this were Claude models (models #10, #11 and #12), which occasionally deviated from the requested format by including a short introductory text. In these instances, the textual outputs were programmatically reformatted to conform to the required JSON structure. It is relevant to note that these three models are the only ones in the evaluation that do not natively support the Structured Output functionality, a factor that likely contributed to their formatting inconsistencies.

Performance Metric

To quantify the models' performance with the given tasks, a robust evaluation metric was required. For this purpose, **Kendall's rank correlation coefficient** ("Kendall's τ", "τ") was selected as the most appropriate measure. Kendall's τ is a non-parametric statistic that measures the ordinal association between two ranked lists. Its methodology is centered on comparing the concordance of all possible pairs of items within the sequences, yielding a score in the interval from -1 (perfect reversal) to +1 (perfect match). The focus on relative, pairwise ordering makes Kendall's τ exceptionally well-suited for a chronological sorting task, as the core challenge lies in correctly establishing which event occurred before another, which is precisely what the metric evaluates.

An alternative metric, the sum of absolute Manhattan distances, was also considered but ultimately deemed less suitable. Its primary drawback is its sensitivity to the magnitude of displacement, which can produce misleading evaluations by heavily penalizing single items that are wildly out of place, while potentially under-penalizing a sequence with numerous smaller, local errors that might represent a poorer overall sort.

Performance by Tasks and Scenarios

The performance of each model, quantified by the Kendall's τ, is detailed in Tables 2 and 3. Table 2 presents the coefficients organized by task (event chain), averaged across all experimental scenarios (prompts). Table 3, in turn, presents the coefficients organized by experimental scenario, averaged across all the tasks. The ranks in both tables were determined by averaging the performance rankings of all the models across individual tasks and scenarios. They largely correspond to the rankings based on average τ, but discrepancies may arise from variation in the scale and distribution of τ values across experiments.

To contextualize these performance metrics, their relationship to pairwise accuracy is critical: within a 10-item sequence, a Kendall's τ of 0.90, 0.80 or 0.50 indicates that approximately 95%, 90% or 75% of the 45 possible pairs are concordantly ordered, respectively.

The aggregated results in Table 2 underscore two principal findings. First, a significant and systematic variation in task difficulty was evident, with Task_1 representing the simplest case and Task_3 the most demanding. This pattern held true for practically all the evaluated models and experimental scenarios. The performance differences indicating different task difficulty were substantial. For Task_1 and the unassisted scenario, the Kendall's τ values for the average, best model, and worst model performance were 0.78, 0.91 and 0.02, respectively. For Task_2, the values were 0.63, 1.00 and 0.16, and for Task_3, they were 0.02, 0.38 and 0.33. These findings clearly establish Task_3 as the most difficult of the three tasks evaluated. Note that a negative Kendall's τ value indicates an inverse correlation between the predicted and true rankings, and a value around zero represents a random ordering. Second, the results show that the more recent versions and models with strong reasoning capabilities (models Grok 4, GPT-5, o3 and o3-pro, and Gemini 2.5 Pro) consistently outperform other models practically across all tasks.

Table 2: Average Performance by Tasks (Kendall's τ)

Rank	Model #	Task_1	Task_2	Task_3	Avg. τ
1	9	0.96	0.98	0.70	0.88
2	2	0.94	0.94	0.56	0.81
3	5	0.96	0.99	0.49	0.81
4	3	0.94	0.93	0.52	0.80
5	6	0.94	0.96	0.52	0.81
6	8	0.93	0.79	0.43	0.72
7	12	0.94	0.70	0.41	0.69
8	20	0.83	0.82	0.50	0.72
9	7	0.84	0.89	0.48	0.74
10	11	0.93	0.67	0.36	0.65
Avg. top 5:		0.95	0.96	0.56	0.82
Avg. all 20:		0.85	0.71	0.36	0.64

The aggregated results in Table 3 underscore three principal findings. First, assisted prompting systematically improved the performance across all models and tasks, which is logical and expected since additional relevant information is provided to the models. Anchoring with known positions in the majority of cases helped the models to better position the remaining articles as well.

Table 3: Average Performance by Scenarios (Kendall's τ)

Rank	Model #	Assist_ No	Assist_ First	Assist_ Last	Assist_ FirstLast	Avg. τ
1	9	0.75	0.88	0.90	0.99	0.88
2	2	0.69	0.88	0.76	0.93	0.81
3	5	0.73	0.84	0.81	0.87	0.81
4	3	0.72	0.87	0.76	0.85	0.80
5	6	0.57	0.93	0.84	0.90	0.81
6	8	0.48	0.81	0.78	0.81	0.72
7	12	0.48	0.66	0.73	0.87	0.69
8	20	0.66	0.75	0.64	0.82	0.72
9	7	0.54	0.73	0.81	0.87	0.74
10	11	0.48	0.64	0.66	0.82	0.65
Avg. top 5:		0.69	0.88	0.81	0.91	0.82
Avg. all 20:		0.47	0.67	0.63	0.79	0.64

Second, the provision of additional information proved more beneficial for the most demanding task (Task_3) than for the less demanding tasks (Task_1 and Task_2). For example, in the Assist_FirstLast scenario, the increase in average τ relative to the unassisted scenario was 0.13 for Task_1, 0.17 for Task_2, and 0.65 for Task_3. This finding follows logically from the models' greater ability to identify the first and/or last article in simpler tasks by themselves: in Task_1, 15 of 20 models correctly identified the first position, while none identified the last position, in Task_2 9 models identified the first position and 4 identified the last position, and in Task_3 no model identified either position correctly.

Third, the provision of additional information disproportionately benefited models that performed poorly in the unassisted scenario. For instance, on Task_3 — the most difficult task with

an average Kendall's τ of only 0.02 in the unassisted scenario — the Assist_First scenario yielded average and maximum performance improvements of 0.46 and 1.07, respectively. For the Assist_Last scenario, the corresponding improvements were 0.27 and 0.80, while for the Assist_FirstLast scenario they were 0.65 and 1.02. The results demonstrate that supplementing less capable models with limited key information can yield significant performance gains at these tasks.

A qualitative examination of the models' reasoning justifications failed to yield systematic insights into their capacity to reconstruct accurate chronological sequences of articles. Although the generated rationales were generally logical and relevant, they frequently omitted crucial contextual information essential for correct chronological reasoning. This observation underscores the challenge that certain timelines may not be uniquely re-constructible due to insufficient contextual information. Furthermore, in some instances, the provided justification could plausibly support an alternative, yet equally valid, timeline. Moreover, this is compounded by the inherent challenge of discerning whether the provided reasoning justifications represent the model's actual inferential process or are merely a result of the post-hoc rationalization.

4 CONCLUSIONS AND FURTHER RESEARCH IDEAS

This research provides insight into the practical application and inherent challenges of utilizing LLMs to sequence news streams in the context of ERM. The selected use cases are based on real-world, business-relevant event chains.

A comparative analysis reveals significant performance disparities among the evaluated models across all tasks and experimental scenarios. Models with superior reasoning capabilities surpassed those with less developed abilities. The varying complexity of the presented tasks further accentuated these performance differences. Also, providing additional anchoring information disproportionately benefited models that performed poorly in the unassisted scenario. Five models, **Grok 4** (xAI), **GPT-5**, **o3** and **o3-pro** (all three OpenAI), and **Gemini 2.5 Pro** (Google), consistently outperformed all other models in practically every task and experiment scenario. The performance level achieved by these models demonstrates their practical utility for real-world ERM applications.

This research has opened several promising areas for **further research**:

(1) Benchmarking LLMs against human experts: A rigorous comparative study should be undertaken in which large LLMs and domain specialists (human experts) perform identical tasks under strictly matched contextual conditions.

(2) Systematically varying model settings to probe "creativity" and reliability: Experiments that modulate the temperature and other model settings can clarify how stochasticity affects task performance and reliability.

(3) Enabling models to request task-critical information: Instead of supplying predefined contextual information—such as the first and/or last article in a sequence—future studies might allow the model to query for the minimal supplementary data it deems most informative. This strategy would approximate an active-learning workflow and might even illuminate new modes for human-LLM collaboration.

(4) Diagnosing mis-ordering errors through reasoning audits: To understand why models fail to reconstruct the correct temporal ordering of news articles, one could extract each model's stated reasoning features for every placement decision, then have human experts or adjudicating LLMs rate their accuracy and relevance. Such audits would expose specific deficits in reasoning and could even inform targeted retraining regimes.

(5) Experimenting with extended or interleaved event chains: Evaluating models on substantially longer sequences—or on mixtures of events drawn from multiple chains—would markedly raise task complexity and furnish a stringent benchmark of temporal-reasoning competence for business use cases.

ACKNOWLEDGMENTS

The authors acknowledge the use of LLMs during various stages of this research. These models provided support in tasks such as idea generation, text processing, prompt engineering, methodological exploration, and language optimization. While the LLMs contributed to enhancing efficiency and refining the presentation of this work, all conceptual frameworks, analyses, and interpretations remain the sole responsibility of the authors.

REFERENCES

[1] Y. Cao et al., 'RiskLabs: Predicting Financial Risk Using Large Language Model Based on Multi-Sources Data', Apr. 11, 2024, arXiv: arXiv:2404.07452. doi: 10.48550/arXiv.2404.07452.

[2] A. Kim, M. Muhn, and V. V. Nikolaev, 'From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI', Jul. 11, 2024, Rochester, NY: 4593660. doi: 10.2139/ssrn.4593660.

[3] T. Li and X. Dai, 'Financial Risk Prediction and Management using Machine Learning and Natural Language Processing', ijacsa, vol. 15, no. 6, 2024, doi: 10.14569/IJACSA.2024.0150623.

[4] Y. Wang, 'Generative AI in Operational Risk Management: Harnessing the Future of Finance', May 17, 2023, Rochester, NY: 4452504. doi: 10.2139/ssrn.4452504.

[5] X. Zhu, H. Jin, J. Li, and Y. Wang, 'Topic-Gpt: A Novel Risk Identification Method Based on Large Language Model', Jul. 04, 2024, Social Science Research Network, Rochester, NY: 4885365. doi: 10.2139/ssrn.4885365.

[6] M. Katamaneni, P. Agrawal, S. Veera, A. K. Sahoo, K. Singh Sidhu, and M. F. Hasan, 'AI-Based Risk Management in Financial Services', in 2024 Second International Conference Computational and Characterization Techniques in Engineering & Sciences (IC3TES), Nov. 2024, pp. 1–5. doi: 10.1109/IC3TES62412.2024.10877497.

[7] X. V. Li and F. S. Passino, 'FinDKG: Dynamic Knowledge Graphs with Large Language Models for Detecting Global Trends in Financial Markets', in Proceedings of the 5th ACM International Conference on AI in Finance, Nov. 2024, pp. 573–581. doi: 10.1145/3677052.3698603.

[8] A. Nygaard et al., 'News Risk Alerting System (NRAS): A Data-Driven LLM Approach to Proactive Credit Risk Monitoring', in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, F. Dernoncourt, D. Preoțiuc-Pietro, and A. Shimorina, Eds., Miami, Florida, US: Association for Computational Linguistics, Nov. 2024, pp. 429–439. doi: 10.18653/v1/2024.emnlp-industry.32.

[9] Z. Xiao, Z. Mai, Z. Xu, Y. Cui, and J. Li, 'Corporate Event Predictions Using Large Language Models', in 2023 10th International Conference on Soft Computing & Machine Intelligence (ISCMI), Nov. 2023, pp. 193–197. doi: 10.1109/ISCMI59957.2023.10458651.

[10] Committee of Sponsoring Organizations of the Treadway Commission (COSO), Enterprise Risk Management—Integrating with Strategy and Performance. Durham, NC: COSO, 2017.

[11] International Organization for Standardization, ISO 31000:2018 – Risk management — Guidelines. Geneva, Switzerland: ISO, 2018.