Counterfactual Causal Inference in Natural Language with Large Language Models

Gaël Gendron[1], Jože M. Rožanec[2], Michael Witbrock[1], Gillian Dobbie[1]

[1]NAOInstitute, University of Auckland, New Zealand [2]Jožef Stefan Institute, Slovenia
Correspondence: gael.gendron@auckland.ac.nz

Abstract

Causal structure discovery methods are commonly applied to structured data where the causal variables are known and where statistical testing can be used to assess the causal relationships. By contrast, recovering a causal structure from unstructured natural language data such as news articles contains numerous challenges due to the absence of known variables or counterfactual data to estimate the causal links. Large Language Models (LLMs) have shown promising results in this direction but also exhibit limitations. This work investigates LLM's abilities to build causal graphs from text documents and perform counterfactual causal inference. We propose an end-to-end causal structure discovery and causal inference method from natural language: we first use an LLM to extract the instantiated causal variables from text data and build a causal graph. We merge causal graphs from multiple data sources to represent the most exhaustive set of causes possible. We then conduct counterfactual inference on the estimated graph. The causal graph conditioning allows reduction of LLM biases and better represents the causal estimands. We demonstrate the applicability of our method on real-world news articles.

1 Introduction

Recovering the causal structure of events described in single or multiple text sources is an important problem for natural language understanding, analysis, and prediction. In particular, causal structure discovery from news articles can help build causal world models that can be used to understand the causal chains behind events, forecast future events, and create robust automated reasoning agents [42, 1]. This problem is not often tackled in natural language processing (NLP) because it requires solving multiple challenges considered open research problems in causality and NLP research. First, the text modality prevents the direct use of traditional structure discovery methods as the set of causal variables is not available and has to be discovered (a problem being recently tackled by the field of causal representation learning) [42]. Second, real-world events have non-trivial structures prone to latent variables and feedback loops, typically excluded from most causal analyses using Direct Acyclic Graphs (DAGs) [2]. Third, causal models provide a means to answer interventional and counterfactual queries, but real-world data prevents the ground truth from being directly accessible. This is the *fundamental problem of causal inference* [32]: only one factual world can be observed. In addition, real-world events are often complex and multi-causal, greatly hindering the possibility of manually annotating a ground truth.

Large Language Models (LLMs) have demonstrated impressive abilities to solve tasks related to these problems, notably for understanding and summarising natural language [7, 4, 5]. Recent work [20] has also shown that LLMs can recover causal structures, although the authors do not apply it to text data. However, this ability is still debated as the LLM's performance can drop significantly when facing unfamiliar settings [19]. Moreover, while being able to perform some causal reasoning tasks

NeurIPS 2024 Workshop on Causality and Large Models (CaLM).

successfully [27], notably on commonsense causal reasoning [22, 52], LLMs have shown limitations on tasks that require robust reasoning, such as arithmetic tasks in unfamiliar settings [47], or abstract reasoning [13]. To explain this behavior difference, [51] advanced that LLMs cannot discover new causal relationships but only recall ones already seen during training. Acknowledging this limitation, alternative approaches have been proposed using LLMs to extract causal relationships from text [41, 38]. However, such approaches do not formally test whether the extracted relationships are causal.

We investigate ways to overcome this restriction and propose an end-to-end causal structure discovery and counterfactual inference method from purely unstructured natural language text data. Our framework is divided into two steps: first, we use an LLM to generate the causal graph associated with a document, i.e., that describes the causal relationships between the events depicted in the text. Optionally, we merge causal graphs from multiple sources using a second LLM. Then, we use the built causal structure to perform an atomic intervention on the sequence of events and infer the consequences in this counterfactual scenario using an LLM (i.e., answer *what if?* questions). We show that our method can effectively extract causal relationships and propose plausible counterfactual worlds from real-world events.

2 Related Work

Causal Structure Discovery with LLMs Text data is often unstructured, high-dimensional, and large-scale. Causal variables may not be directly accessible from the text, and the causal relationships are typically vague and rare, with semantic ambiguity complicating analysis. These challenges greatly hinder the usability of traditional causal structure discovery methods and motivate using LLMs for causal structure discovery [22, 25]. [22] have achieved promising results when using LLMs to infer the causal direction between two variables. Nevertheless, there is some evidence that LLMs, in many cases, repeat embedded causal knowledge [51] and are susceptible to inferring causal relations from the order of two entities mentioned in a text [21]. [20] attempts to recover the full causal structure using a breadth-first search on a set of text variables. [17] investigates whether LLMs can identify the cause and the effect between two natural language sentences. This line of work uses the LLM's inner knowledge to discover causal relationships between data points. However, recent work highlighted that LLMs do not conduct proper causal reasoning and mainly rely on domain knowledge and correlations [51, 19]. This approach differs from another line of work that uses the LLM as an information retrieval engine to extract causal relationships explicitly present in the data. For instance, [15] uses LLMs to extract causal relationships from medical texts. Our approach combines both worlds as we use an LLM to retrieve causal relationships from text and then perform causal inference on the extracted model to assess the quality of the causal model. NATURAL [8] is another method developed concurrently to our work using LLMs to perform causal inference. However, the method is based on the Potential Outcome Framework (POF) and computes the Average Treatment Effect (ATE) of a variable (i.e., outcome) under an intervention (i.e., a treatment). By contrast, our method is based on Structural Causal Models (SCMs) and includes a causal structure discovery step to represent complex causal relationships between observed variables. Our work also focuses on counterfactuals, while this method is suited to answer interventional queries.

Strategic Foresight Strategic Foresight aims to provide a structured approach to gathering information regarding plausible future scenarios and adequately preparing for change. It provides expert insights regarding trends and emerging issues that can be considered for strategic planning and policy-making. As such, it is being increasingly adopted in the public and private sectors [6, 37]. Among the most frequently used methods, we find scenario planning [10], which aims to foresee relevant scenarios based on trends and factors of influence to understand better how actions can influence the future [46]. While the value of artificial intelligence for strategic foresight has been recognized, much of the work is still not automated [35, 3]. Scientific literature reports on using artificial intelligence for information scanning and analysis [31, 3], to identify weak signals and trends [14], and extract actions and outcomes that can be mapped to causal decision diagrams [33]. More recently, authors have proposed architectures that could automate strategic foresight. Nevertheless, the proposed architecture only considered a signal assessment module without explicit reference to testing causality among the extracted graph relationships [39, 40]. We aim to bridge this gap by identifying, extracting, and testing causal relationships reported in media news to construct a graph of causal relationships and use such a graph to build plausible future scenarios, providing expert insights for strategic planning and policy-making.

3 Counterfactual Inference with Large Language Models

This section describes our proposed method for causal structure discovery and causal inference from text. We build a modified version of the Structural Causal Model (SCM) [32] to represent the described causal mechanisms. Following the SCM framework, we consider that the mechanisms can be represented as Directed Acyclic Graphs (DAGs). For every document D, we construct a DAG $\mathcal{G} = \langle \mathbf{U}, \mathbf{V}, \mathbf{E} \rangle$ where the observed nodes $v \in V$ correspond to general depictions of the events in the text. Two nodes $(V_i, V_j) \in \mathbf{V} \times \mathbf{V}$ are connected by an edge $E_{ij} = (V_i, V_j) \in \mathbf{E}$ if a causal relationship between them is explicitly mentioned in the text data (according to an LLM). We also represent possible exogenous factors $U \in \mathbf{U}$ describing unobserved events having a causal influence on the observations. Nodes and edges also have features. Nodes correspond to causal variables and have a *domain* established during extraction and a *current value* from this domain. Confounders U are not assigned values because they are not observed. As we aim to use an LLM for inference, we also keep attributes in plain natural language: a *description* of the variable and additional *contextual information*. Edges also have a *description* attribute. For example, given this sentence: "The airlines companies have seen their revenues diminishing due to travel restrictions.", we can extract the following causal variables S and T, and their relationship: *travel restrictions (S)* $\rightarrow$ *airlines revenues (T)*. Their domains can be, e.g., a boolean for S and a fixed set of categories for T (since we do not have access to the numerical values of the revenues). Their current values are written $S = s_o$ and $T = t_o$. During counterfactual inference, we intervene to modify these values while the other attributes remain unchanged. For performance, we add contextual information to each node, i.e., background knowledge extracted from the text. In this example, the contextual information could be the country where the event takes place. The edge description can be, e.g., "travel restrictions diminish airlines companies revenues".

SCMs typically have a set of mapping functions $\mathcal{F}$ to infer the value of a variable V given its parents $\mathbf{pa}(\cdot)$ (e.g., $V \leftarrow f_V(\mathbf{pa}(V))$). Instead of using a set of predefined functions, we perform causal inference using an LLM. We also use it to compute the prior probability distribution of the confounders U. We discuss our method and the implications of using an LLM for inference in Section 3.2. We note a full causal model containing the causal graph $\mathcal{G}$ and all these attributes as $\mathcal{M} = \langle \mathcal{G}, \text{LLM} \rangle$. The instantiated model $\mathcal{M}(D)$ describes the model with the values of each variable extracted from the document D. We further note an intervention $do(X = x)$ in this model as $\mathcal{M}_{X=x}(D)$ or $\mathcal{M}(D, do(X = x))$.

Our proposed method is divided into four stages. First, we extract the causal graph from a text document. If multiple documents are provided, we merge their respective causal graphs together. Then, we compute counterfactual worlds from the resulting causal graph. We use these counterfactuals to self-evaluate the causal graph. Section 3.1 describes the causal graph extraction step. Section 3.2 describes the counterfactual inference step. Evaluation is described in Sections **??** and 4. Figure 1 illustrates the complete pipeline.

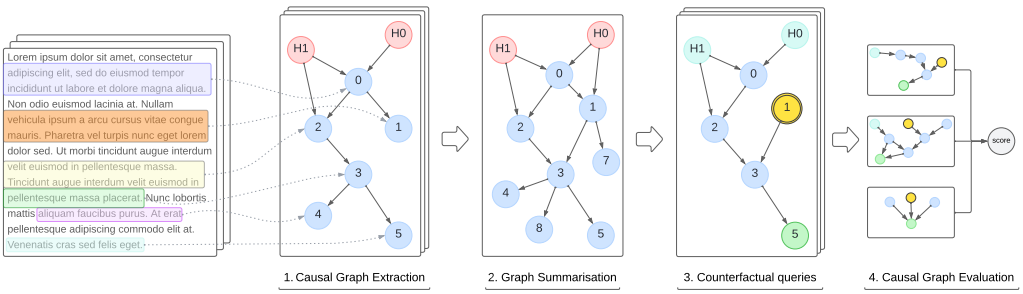


Figure 1: Overview of the proposed framework. (1) An LLM extracts causal variables and their corresponding causal relationships from the input text. (2) Multiple graphs are generated and merged into a single graph if multiple text snippets are given in input. (3) The resulting causal graph is edited using ablation, intervention, and prediction steps to build counterfactual instantiations. The LLM performs inference given the variables' parent values. (4) The LLM self-evaluates the original and counterfactual graphs.

3.1 Causal Structure Discovery from Natural Language

We prompt an LLM to read an input document and return its associated causal graph. The expected graph should contain the full set of observed causal variables, their relationships, and their attributes as described in the introduction of Section 3. We also ask the LLM to estimate the hidden variables affecting the observed events and how they are connected. Estimated hidden variables have the same attributes as the observed ones but do not have an observed value. Other works have studied ways to improve the quality of the generated causal graphs, e.g., via breadth-first search [20]. However, to keep our pipeline efficient, we only consider a single forward pass with chain-of-thought prompting [45]. We find that this simple choice is sufficient for our purpose. This is not surprising as the causal relationships to be found are explicitly described in the data and LLMs have been very successful at information retrieval tasks [4, 5, 36]. We prompt the LLM to return its answer in JSON format. Still, it may not always provide an extractable response. To alleviate this issue, we allow the LLM to refine its answer several times if it cannot be parsed automatically. If multiple documents are provided, we merge the causal graphs together. We describe this optional step in Appendix A.

3.2 Counterfactual Causal Inference

Autoregressive LLMs are inference engines computing the conditional probability of an output token Y_0 given an input context $\mathbf{C}$: $P(Y_0|\mathbf{C})$. When used for generation, they construct an output sequence $\mathbf{Y} = [Y_0, \ldots, Y_N]$ by computing $P(Y_i|Y_{i-1}, \ldots, Y_0, \mathbf{C})$ iteratively. Due to their extensive training on a massive amount of data, LLMs are good estimators of $P(Y_0|\mathbf{C})$ [4]. However, LLMs are also prone to hallucinations when providing long answers: they deviate from the instructions or state false information [18]. Indeed, estimating the true conditional distribution of the full output $P(\mathbf{Y}|\mathbf{C})$ is more challenging, especially when $\mathbf{Y}$ is long as it requires building a probability tree considering all possible values for the intermediate Y_i. This tree can be approximated using beam search or heuristic-guided tree search algorithms [49, 44]. However, their performance is still dependent on the output length. We alleviate this problem by conditioning the inference query on the causal parents of the output, i.e. instead of providing the complete context $\mathbf{C}$ as an input of the LLM, we use the much smaller subset $\mathbf{pa}(\mathbf{Y}) \subset \mathbf{C}$. Assuming knowledge of the causal graph, this choice can greatly reduce the size of the context window and mitigate hallucination. In the rest of this section, we assume that the LLM can provide a close estimate of the true conditional distribution $P(\mathbf{Y}|\mathbf{pa}(\mathbf{Y}))$. We challenge this assumption in our experiments. We investigate LLMs' causal inference abilities in counterfactual settings given the estimated causal model $\mathcal{M}$. Counterfactual queries answer the question: "How would variable Y change if we had $X = x$ instead of $X = x'$?". This question can be answered by performing *abduction*, *intervention* and *prediction*. The abduction step estimates the values of the exogenous factors U from the observed quantities: $P(U|x', y')$. The intervention step edits the causal graph with the $do(X = x)$ operation. The prediction step computes the remaining variables from their parent values: $P(Y|\mathbf{pa}(Y))$. The corresponding quantity is expressed as follows:

$$P(Y|do(x), x', y') = \sum_{u \in U} P(Y|do(x), u)P(u|x', y') \tag{1}$$

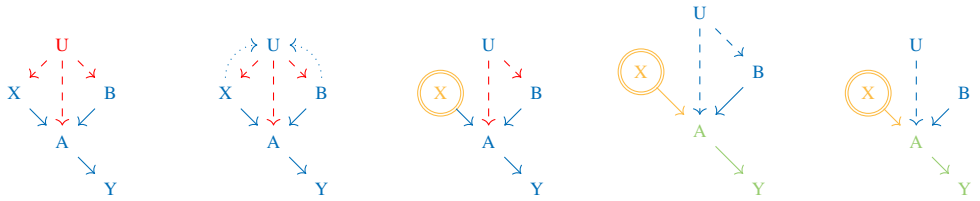
Figure 2 illustrates these steps. To be efficient, we approximate some of them. We sample a single $u \sim P(U|\mathbf{ch}(U))$ using the LLM. $\mathbf{ch}(\cdot)$ represents the children of U. We perform the abduction and prediction steps only on the variables affected by the intervention, as shown in Figure 2e where we use the LLM as described above to compute $P(A|X, B, U)$ and $P(Y|A)$.

4 Experiments on Real-World Usecase

4.1 Experimental Setup

We extract 5,486 media events related to the "Price of Oil" from EventRegistry [24]. These events, spanning the first quarters of 2015, 2020, 2022, and 2023, were selected based on geopolitical events highlighted by the U.S. Energy Information Administration and the Russo-Ukrainian War [1]. We perform experiments using LLaMA-3.1 [9] and GPT-4o [28, 29]. We provide early results on

[1]The geopolitical events were highlighted in the following report, last accessed on August 25th 2023: `https://www.eia.gov/finance/markets/crudeoil/spot_prices.php`.



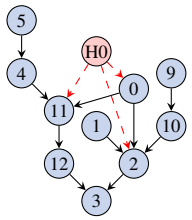
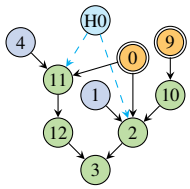
(a) Factual graph. (b) Abduction step. (c) Intervention step. (d) Prediction step. (e) Optimised version.

Figure 2: Counterfactual inference steps. (2a) The original causal graph. (2b) We estimate the possible values of the exogenous factors, here U, from the observations. (2c) We perform the $do(X = x)$ operation. (2d) We predict the values of the remaining variables given their parent causes. Here, X and U are known. B, A and Y should be predicted. (2e) However, to maintain efficiency, we consider a single possible value per exogenous factor and re-compute only the variables affected by the intervention: here B is unaffected and not re-computed.

real-world news documents. Due to the unavailability of the ground truth, we focus our experiments on a handful of documents that we manually verify.

4.2 Counterfactual Inference

Table 1: Example of counterfactual inference performed by the Counterfactual-CI model on real-world data. The first row shows the extracted factual graph and the counterfactual graph under interventions (in yellow) *do(0='low')* and *do(9='False')*. The exogenous factor is in red. After abduction, a value is assigned to it (illustrated in blue). Variables inferred during the prediction step are shown in green. The bottom rows show the values of the factual and counterfactual worlds.

Factual Graph	Intervened Graph
	

	Variable Descriptions	Factual Values	Counterfactual Values GPT-4o	Counterfactual Values LLaMA-3.1
0	Severity of COVID-19 pandemic (range element)	severe	low (from do operation)	
1	Severity of oil price war (range element)	severe		
2	Bursa Malaysia downtrend magnitude (int)	29%	20	high
3	FBM KLCI index value (float)	1,280.63	1580	1500.00
4	Selling pressure on stocks (range element)	high		
5	Investors moving into cash (bool)	true		
9	Malaysia's change of coalition government (bool)	true	false (from do operation)	
10	Downside risks to corporate earnings (range element)	high	low	low
11	Travel restrictions imposed worldwide (range element)	severe	none	none
12	Condition of oil & gas and airlines sectors (range element)	bad	good	good
h0	Potential end of COVID-19 pandemic (bool)	none	false	true

Table 1 provides an example of results obtained with our method for a model $\mathcal{M}(D)$ extracted from a document D and under interventions $\mathcal{M}_{0='low',9='False'}(D)$. The input text and the explanations given by the LLMs during inference are given in Appendix C. The document D describes how the COVID-19 pandemic and the rise in oil prices affect Malaysia's economy. We perform two interventions on the graph and build a counterfactual world where the severity of the pandemic is low and a change of governement has not happened. From these interventions, the model updates the rest of the variables and predicts a better economical situation. We emphasise that these results should be

taken as an illustrative example only. We also observe that the two models return different values for the exogenous factor but reach the same conclusions.

We also ask the LLM to self-evaluate its generated graphs. We generate four causal graphs from a single document and, for each graph, we automatically build six counterfactual queries. We summarise each graph into text format and prompt the LLM to return a plausibility score for the chain of events and a confidence score for its prediction (prompts are given in Appendix B.4). We show the results in Table 2. The LLMs generally provide higher scores to the factual graph than to the counterfactuals. This is not surprising as the former are extracted from real data, closer to the LLMs' training distribution. However, they still give high scores to the built counterfactuals. GPT-4o is also more consistent on the factual graphs, with a lower standard deviation, highlighting consistency between graph generations. LLaMA-3.1 shows a high standard deviation for all results.

Table 2: Self-evaluation and confidence provided by the LLMs on the plausibility of the described set of events and their causal relationships for the document described in Appendix B. The average of three end-to-end runs is shown.

	Self-Evaluation		Self-Confidence	
	Factual	Counterfactuals	Factual	Counterfactuals
Counterfactual-CI-GPT-4o	0.850 ± 0.000	0.739 ± 0.129	0.863 ± 0.022	0.762 ± 0.155
-LLaMA-3.1	0.563 ± 0.330	0.256 ± 0.340	0.600 ± 0.346	0.367 ± 0.400

5 Limitations

LLMs provide free-text answers that can be different from the format provided in the instructions. This poses challenges for building an end-to-end pipeline that requires using LLMs at multiple stages, as the responses can be difficult to parse automatically, and errors can accumulate. In our future work, we will include fine-tuning in our pipeline to mitigate this issue. Due to the high cost of running LLMs or accessing them through APIs, we only tested our method on short text snippets. We intend to apply it to larger amounts of data in the future. We also only consider DAG structures, whereas real-world events can contain feedback loops. We will integrate them into our future work. As the LLM discovers the causal structure, errors can be present, and confounders can be omitted. The counterfactual results depend on the causal graph to be accurate. In addition, our current approach proxies ground-truth counterfactual data by retrieving such values from LLMs. Our future work will focus on verifying the accuracy of the intermediate steps and building ground-truth counterfactual data.

6 Conclusion and Broader Impact

We propose an end-to-end method for conducting causal structure discovery and counterfactual causal inference from unstructured natural language. We demonstrate the applicability of our method on real-world news events, showcasing the LLMs' abilities to perform causal discovery and inference not as a standalone model but as a part of a larger framework.

This research is still in its infancy but has shown promising results for computing causal inference and leveraging LLMs. As the inference process is divided into several independent computations of causal variables given their parents, this configuration provides inherent interpretability and allows auditing an LLM's answer. Our future work will follow two lines of research. On one hand, as LLMs have been shown to perform better by using refinement techniques [47, 26, 50, 34], our counterfactual self-evaluation method could be used to conduct *Counterfactual Self-Learning*: refine LLMs' answers to improve them and teach LLMs to reason more causally by providing them with counterfactual data [1]. On the other hand, we aim to create a framework to build and test counterfactuals based on real values and validate whether LLM-extracted causal relationships hold. Doing so would avoid LLM hallucinations and lead to robust causal reasoning, where every piece of information could be traced back to evidence from the real world. We expect that such a framework will have extensive applications in many domains, extending the use of causal reasoning to any domain where text is available. In particular, we expect that could lead to a greater automation of strategic foresight, democratizing and enabling a wider use of it to enhance decision-making at all societal levels.

References

[1] Elias Bareinboim et al. "On Pearl's Hierarchy and the Foundations of Causal Inference". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. Ed. by Hector Geffner, Rina Dechter, and Joseph Y. Halpern. Vol. 36. ACM Books. ACM, 2022, pp. 507–556. DOI: 10.1145/3501714.3501743. URL: https://doi.org/10.1145/3501714.3501743.

[2] Stephan Bongers et al. "Foundations of structural causal models with cycles and latent variables". In: *The Annals of Statistics* 49.5 (2021), pp. 2885–2915. DOI: 10.1214/21-AOS2064. URL: https://doi.org/10.1214/21-AOS2064.

[3] Patrick Brandtner and Marius Mates. "Artificial Intelligence in Strategic Foresight–Current Practices and Future Application Potentials: Current Practices and Future Application Potentials". In: *The 2021 12th International Conference on E-business, Management and Economics*. 2021, pp. 75–81.

[4] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[5] Sébastien Bubeck et al. "Sparks of Artificial General Intelligence: Early experiments with GPT-4". In: *CoRR* abs/2303.12712 (2023). DOI: 10.48550/arXiv.2303.12712. arXiv: 2303.12712. URL: https://doi.org/10.48550/arXiv.2303.12712.

[6] George Burt and Anup Karath Nair. "Rigidities of imagination in scenario planning: Strategic foresight through 'Unlearning'". In: *Technological Forecasting and Social Change* 153 (2020), p. 119927.

[7] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/V1/N19-1423. URL: https://doi.org/10.18653/v1/n19-1423.

[8] Nikita Dhawan et al. "End-To-End Causal Effect Estimation from Unstructured Natural Language Data". In: *CoRR* abs/2407.07018 (2024). DOI: 10.48550/ARXIV.2407.07018. arXiv: 2407.07018. URL: https://doi.org/10.48550/arXiv.2407.07018.

[9] Abhimanyu Dubey et al. "The Llama 3 Herd of Models". In: *CoRR* abs/2407.21783 (2024). DOI: 10.48550/ARXIV.2407.21783. arXiv: 2407.21783. URL: https://doi.org/10.48550/arXiv.2407.21783.

[10] Ashkan Ebadi, Alain Auger, and Yvan Gauthier. "Detecting emerging technologies and their evolution using deep learning and weak signal analysis". In: *Journal of Informetrics* 16.4 (2022), p. 101344.

[11] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[12] Kenneth D. Forbus. "Analogical Abduction and Prediction: Their Impact on Deception". In: *2015 AAAI Fall Symposia, Arlington, Virginia, USA, November 12-14, 2015*. AAAI Press, 2015, pp. 15–20. URL: http://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11660.

[13] Gaël Gendron et al. "Large Language Models Are Not Strong Abstract Reasoners". In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. Ed. by Kate Larson. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2024, pp. 6270–6278. DOI: 10.24963/ijcai.2024/693. URL: https://doi.org/10.24963/ijcai.2024/693.

[14] Amber Geurts et al. "New perspectives for data-supported foresight: The hybrid AI-expert approach". In: *Futures & Foresight Science* 4.1 (2022), e99.

[15] Seethalakshmi Gopalakrishnan, Luciana Garbayo, and Wlodek Zadrozny. "Causality extraction from medical text using Large Language Models (LLMs)". In: *arXiv preprint arXiv:2407.10020* (2024).

[16] Luigi Gresele et al. "Causal Inference Through the Structural Causal Marginal Problem". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 7793–7824. URL: https://proceedings.mlr.press/v162/gresele22a.html.

[17] Marius Hobbhahn, Tom Lieberum, and David Seiler. "Investigating causal understanding in LLMs". In: *NeurIPS ML Safety Workshop*. 2022.

[18] Lei Huang et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *CoRR* abs/2311.05232 (2023). DOI: 10.48550/ARXIV.2311.05232. arXiv: 2311.05232. URL: https://doi.org/10.48550/arXiv.2311.05232.

[19] Zhijing Jin et al. "Can Large Language Models Infer Causation from Correlation?" In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=vqIH0ObdqL.

[20] Thomas Jiralerspong et al. "Efficient Causal Graph Discovery Using Large Language Models". In: *CoRR* abs/2402.01207 (2024). DOI: 10.48550/ARXIV.2402.01207. arXiv: 2402.01207. URL: https://doi.org/10.48550/arXiv.2402.01207.

[21] Nitish Joshi et al. "LLMs Are Prone to Fallacies in Causal Inference". In: *arXiv preprint arXiv:2406.12158* (2024).

[22] Emre Kiciman et al. "Causal Reasoning and Large Language Models: Opening a New Frontier for Causality". In: *CoRR* abs/2305.00050 (2023). DOI: 10.48550/arXiv.2305.00050. arXiv: 2305.00050. URL: https://doi.org/10.48550/arXiv.2305.00050.

[23] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[24] Gregor Leban et al. "Event registry: learning about world events from news". In: *Proceedings of the 23rd International Conference on World Wide Web*. 2014, pp. 107–110.

[25] Jing Ma. "Causal Inference with Large Language Model: A Survey". In: *arXiv preprint arXiv:2409.09822* (2024).

[26] Aman Madaan et al. "Self-Refine: Iterative Refinement with Self-Feedback". In: *CoRR* abs/2303.17651 (2023). DOI: 10.48550/arXiv.2303.17651. arXiv: 2303.17651. URL: https://doi.org/10.48550/arXiv.2303.17651.

[27] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. "Causal Transformer for Estimating Counterfactual Outcomes". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 15293–15329. URL: https://proceedings.mlr.press/v162/melnychuk22a.html.

[28] OpenAI. "GPT-4 Technical Report". In: *CoRR* abs/2303.08774 (2023). DOI: 10.48550/ARXIV.2303.08774. arXiv: 2303.08774. URL: https://doi.org/10.48550/arXiv.2303.08774.

[29] OpenAI. *Hello GPT-4o*. Accessed: 2024-09-23. 2024. URL: https://openai.com/index/hello-gpt-4o/.

[30] Matías Osta-Vélez and Peter Gärdenfors. "Analogy as a search procedure: a dimensional view". In: *Journal of Experimental & Theoretical Artificial Intelligence* 0.0 (2022), pp. 1–20. DOI: 10.1080/0952813X.2022.2125081. eprint: https://doi.org/10.1080/0952813X.2022.2125081. URL: https://doi.org/10.1080/0952813X.2022.2125081.

[31] Nathan H Parrish et al. "Crystal cube: Multidisciplinary approach to disruptive events prediction". In: *Advances in Human Factors, Business Management and Society: Proceedings of the AHFE 2018 International Conference on Human Factors, Business Management and Society, July 21-25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9*. Springer. 2019, pp. 571–581.

[32] Judea Pearl. *Causality*. 2nd ed. Cambridge university press, 2009. DOI: 10.1017/CBO9780511803161.

[33] Lorien Pratt, Christophe Bisson, and Thierry Warin. "Bringing advanced technology to strategic decision-making: The Decision Intelligence/Data Science (DI/DS) Integration framework". In: *Futures* 152 (2023), p. 103217.

[34] Linlu Qiu et al. "Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement". In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=bNt7oajl2a.

[35] Norbert Reez. "Foresight-Based Leadership. Decision-Making in a Growing AI Environment". In: *International Security Management: New Solutions to Complexity*. Springer, 2020, pp. 323–341.

[36] Machel Reid et al. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". In: *CoRR* abs/2403.05530 (2024). DOI: 10.48550/ARXIV.2403.05530. arXiv: 2403.05530. URL: https://doi.org/10.48550/arXiv.2403.05530.

[37] Aaron B Rosa, Niklas Gudowsky, and Petteri Repo. "Sensemaking and lens-shaping: Identifying citizen contributions to foresight through comparative topic modelling". In: *Futures* 129 (2021), p. 102733.

[38] Jose Rozanec et al. "Back to the Future: predicting causal relationships influencing oil prices". In: *The Second Tiny Papers Track at ICLR 2024*.

[39] Joze Rozanec et al. "AI, what does the future hold for US? Automating strategic foresight". In: *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering*. 2023, pp. 247–248.

[40] Jože M Rožanec et al. "AI-based Strategic Foresight for Environment Protection". In: *Symposium on AI, Data and Digitalization (SAIDD 2023)*. Vol. 7. 2023.

[41] Jože M Rožanec et al. "Building A Causality Graph For Strategic Foresight". In: (2023).

[42] Bernhard Schölkopf et al. "Toward Causal Representation Learning". In: *Proc. IEEE* 109.5 (2021), pp. 612–634. DOI: 10.1109/JPROC.2021.3058954. URL: https://doi.org/10.1109/JPROC.2021.3058954.

[43] Petar Velickovic et al. "Graph Attention Networks". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=rJXMpikCZ.

[44] Ziyu Wan et al. "AlphaZero-Like Tree-Search can Guide Large Language Model Decoding and Training". In: *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=C4OpREezgj.

[45] Jason Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *NeurIPS*. 2022. URL: http://papers.nips.cc/paper%5C_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[46] Angela Wilkinson. "Strategic foresight primer". In: *European Political Strategy Centre* (2017).

[47] Zhaofeng Wu et al. "Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks". In: *CoRR* abs/2307.02477 (2023). DOI: 10.48550/arXiv.2307.02477. arXiv: 2307.02477. URL: https://doi.org/10.48550/arXiv.2307.02477.

[48] Keyulu Xu et al. "How Powerful are Graph Neural Networks?" In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=ryGs6iA5Km.

[49] Shunyu Yao et al. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models". In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh et al. 2023. URL: http://papers.nips.cc/paper%5C_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

[50] Weizhe Yuan et al. "Self-Rewarding Language Models". In: *CoRR* abs/2401.10020 (2024). DOI: 10.48550/ARXIV.2401.10020. arXiv: 2401.10020. URL: https://doi.org/10.48550/arXiv.2401.10020.

[51] Matej Zecevic et al. "Causal Parrots: Large Language Models May Talk Causality But Are Not Causal". In: *Trans. Mach. Learn. Res.* 2023 (2023). URL: https://openreview.net/forum?id=tv46tCzs83.

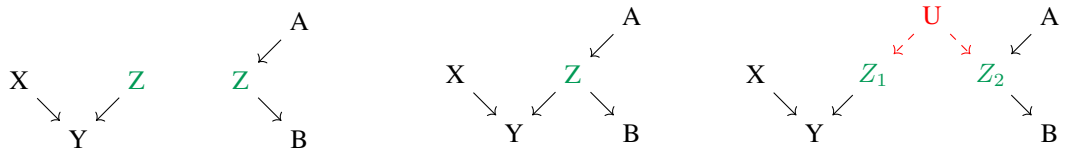
[52] Jiayao Zhang et al. "ROCK: Causal Inference Principles for Reasoning about Commonsense Causality". In: *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 26750–26771. URL: https://proceedings.mlr.press/v162/zhang22am.html.

A Graph Summarisation

When considering several documents, we build their corresponding causal graphs independently and then attempt to merge them. The problem of combining multiple causal graphs together is known as the *Structural Causal Marginal problem* [16]. This is a challenging problem that we approach from two different angles. Considering two causal models $\mathcal{M}_1$ and $\mathcal{M}_2$, we first generate embeddings for every node of the two graphs by extracting their representation generated by the last hidden layer of the LLM. The prompt of a node is a concatenation of all its attributes. We omit its current value because we want to generate a representation of the variable and not the current instance. Motivated by the success of Graph Neural Networks (GNNs) at aggregating neighborhood information in graphs [23, 43, 48], we also add to he prompt the attributes of its immediate neighbours and how they are connected via the edge attributes. Then, we use a clustering algorithm to find nodes that share similar latent representations. We select DBSCAN [11]. Since it is a density-based algorithm, it allows use to restrict the sparsity of the clusters and cluster together nodes with only very close representations.

After extracting similar nodes, we consider two ways to combine graphs: *summarisation* and *analogy*. Summarisation implies considering similar nodes as a single node in the merged graph, inheriting the edges of all the nodes in the set. This approach is straightforward and allows reducing the number of nodes as the graph grows. However, assessing if two causal variables can be merged is a challenging problem. In particular, in the absence of a lot of observations (one text only shows one observation per variable), the merged causal graph can be easily falsified [16].

Analogy merging is inspired by the research on analogical reasoning [30, 12]. We view similarity between nodes representations as an indication that analog mechanisms are causing them. To represent an analogy, we do not modify the existing graphs but add a common unobserved ancestor between similar nodes to integrate the similarity information from the clustering process without making assumptions regarding the nature of their similarity. The merged graphs can share information via backdoor paths. Unlike summarisation, analogy does not remove nodes but adds more. However, this method does not introduce ways to falsify the graphs and preserve the mechanisms of the initial graphs. The two approaches are illustrated in Figure 3.



(a) Model $\mathcal{M}_1$ (b) Model $\mathcal{M}_2$ (c) $\mathcal{M}_{12}^{(\text{Summarisation})}$ (d) $\mathcal{M}_{12}^{(\text{Analogy})}$

Figure 3: Illustration of graph merging process for two models $\mathcal{M}_1$ and $\mathcal{M}_2$. We assume that the common variable Z is the same in the two graphs and should be combined. Figure 3c shows the merged graph using summarisation: Z is shared by the two mechanisms. Figure 3d shows the merged graph using analogy: the variables Z_1 and Z_2 from $\mathcal{M}_1$ and $\mathcal{M}_2$ remain separated but share a common ancestor.

B Prompts Used

B.1 Causal Structure Discovery

Here is the system prompt used for causal discovery:

```
Your task is to summarise a text into a JSON dictionary of instantiated causal variables and the causal
      relationships between them.
Variables should be as atomic and detailed as possible. Causal relationships should describe how the value
       of the first variable affects the value of the second.
One sentence usually describes two or more variables and connects them. For each variable, the following
      questions should be answered:
'What are the causes of this variable's value? Is it fully explained by the available information or are
      some causes missing?'
If some causes seem to be missing, create new (hidden) variables.
Hidden variables represent missing information to fully explain the value of one or more observed
      variables.
They cannot have incoming edges. Identify the major and minor variables and how they are connected.
```

```
Add the missing unknown variables when necessary. Follow carefully the instructions and write down your
      answer using only the given JSON format very strictly.
The format is as follows:
{
  "observed_nodes": [
    {
      "node_id": (str) "0",
      "description": (str) "<high-level short atomic description of causal variable 0>",
      "type": (str) "<variable type: e.g. bool, int, set element, range element>",
      "values": (str) "<set of possible values, if applicable>",
      "current_value": (str) "<current value>",
      "context": (str) "<contextual information type> : <value of the contextual information linked to the
            current instance>"
    },
    ...
  ],
  "hidden_nodes": [
    {
      "node_id": (str) "h0",
      "description": (str) "<high-level short atomic description of the hidden causal variable>",
      "type": (str) "<variable type: e.g. bool, int, set element, range element>",
      "values": (str) "<set of possible values, if applicable>",
      "current_value": (str) "", # This field is left empty because the current value of the variable is
            unknown since the variable is hidden
      "context": (str) "<contextual information type> : <value of the contextual information linked to the
            current instance>"
    },
    ...
  ],
  "observed_edges": [
    {
      "source_node_id": (str) "0",
      "target_node_id": (str) "1",
      "description": (str) "<high-level short atomic description of the causal relationship from variable 0
             to 1>",
      "details": (str) "<detailed explanation of how the value of variable 0 affects the value of variable
            1 in the text>"
    },
    ...
  ],
  "hidden_edges": [
    {
      "source_node_id": (str) "h0",
      "target_node_id": (str) "1",
      "description": (str) "<high-level short atomic description of the causal relationship from hidden
            variable 0 to 1>",
      "details": (str) "<detailed explanation of how the value of hidden variable 0 affects the value of
            variable 1 in the text>"
    },
    ...
  ]
}
```

Here is the parameterised user prompt, specific for each instance. The `{text}` parameter is replaced by the input document.

```
Here is the input text:
‘‘‘
{text}
‘‘‘
```

B.2 Graph Summarisation

When conducting graph summarisation, we do not use the LLM as a generative model but as an embedding model. We only provide node information using the following format. Texts in curly brackets represent variable parameters.

```
description: {description}, type: {type}, values: {values}, context: {context}
```

When adding neighbour information, we concatenate the initial representation with the folowing neighbour representation:

```
neighbour at distance {rank} from node: description: {description}, type: {type}, values: {values},
      context: {context}
```

B.3 Counterfactual Inference

During inference, we provide parent variables to the LLM and prompt it to estimate the value of the target variable. Here is the system prompt:

```
Your task is to predict the value of the target variable given its description, type, possible values, and
      context, and the attributes and values of its parent causes and the relationships connecting them.
The value of the target variable is fully determined by its direct list of causes. Reason step-by-step.
     Start by describing the attributes of the target variable and explain in your own words its
     relationships with its parent causes, how the variables are linked, and how their values cause the
     value of the target. Then, predict the value of the target variable. Provide a confidence score as a
     float between 0 and 1. Follow strictly the provided format.
```

The user prompt provides information about the target variable. The format is as follows:

```
The target variable has the following attributes: {node attributes}.
It is caused by the following variables:
```

The list of parent is then provided. The ith parent variable is described as follows:

```
{i}. {parent attributes}. Its value is {parent value}. Its causal relationship with the target is
     described as follows: {edge attributes}
```

We generate the value of the intervened variable using an LLM. Here is the system prompt provided for this task:

```
Your task is to interpret the attributes of a variable and propose an alternative/counterfactual
     instantiation different from its current value. The variable is described by its description, type,
     possible values, current value, and context. The counterfactual value should be a plausible
     alternative instantiation of the variable given the context, type, description, and possible values.
     Reason step-by-step. Start by describing the attributes of the variable and explain in your own words
      the reasons for the choice of the counterfactual value. Then, state the factual value and propose
     the new counterfactual value. Provide a confidence score as a float between 0 and 1. Follow strictly
     the provided format.
```

Here is the user prompt:

```
The variable has the following attributes: description: {description}, type: {type}, possible values: {
     values}, context: {context}. The current value is {current_value}. Propose a counterfactual value.
```

B.4 Evaluation

The evaluation of the causal factual and counterfactual models is performed by an LLM using the following prompts. Here is the system prompt:

```
Your task is to evaluate the plausibility of a set of events linked by causal relationships. The events
     are described by a high-level description and a value. The events are linked by causal relationships.
      The causal relationships are described by a high-level description. The overall plausibility of the
     set of events corresponds to the factorization of the plausibility of each event's occurrence given
     its causes. Reason step-by-step. Start by describing the events and the causal relationships. Explain
      in your own words the reasons for the plausibility of each event. Finally, provide an overall score
     for the plausibility of the sequence of events. Give an explanation describing your reasoning.
     Provide an overall confidence score as a float between 0 and 1. Follow strictly the provided format.
```

The user prompt describes the events in the topological order of the causal graph. Before each event, the causal relationships with its parents is also described. Here is an example:

```
The causal graph is composed of the following events:
({parent rank} -> {target rank}) {edge description}.
{target rank}. {target description}. The value is {node current_value}
```

C Supplement to the Experiments

In this section, we describe in more details the counterfactual causal inference example from Table 1. The document from which is extracted the causal graph is shown below:

```
KUALA LUMPUR: Bursa Malaysia downtrend could be far from over as there is always more room to decline
      depending on the severity of COVID-19 pandemic and oil price war, said economists after key benchmark
       FBM KLCI took another beating in the early trading session yesterday.
To what may be seen as continued selling pressure from last week's Friday the 13th, the FTSE Bursa
      Malaysia KLCI (FBM KLCI) lost 44.99 points to 1,299.76 at 9.10am yesterday, compared with Friday's
      close of 1,344.75, after opening 25.38 points lower at 1,319.37 yesterday morning.
At market closing yesterday, the FBM KLCI closed at 4.77 per cent lower to 1,280.63 points with turnover
      at 4.473 billion shares valued at RM3.687 billion.
Bank Islam chief economist Dr Mohd Afzanizam Abdul Rashid said if history is of any guide, the FBM KLCI
      has fallen sharply between January 11, 2008 (1,516.22 points) and October 29, 2009 (829.41 points).
He said during the Asia Financial Crisis in 1997/1998, the FBM KLCI was down massively by 79.2 per cent
      between February 25, 1997 (1,265.01 points) and September 1, 1998 (262.7 points).
"For now, the FBM KLCI touches its peak at 1,895.18 points on April 19, 2018 and has plunged by 29.0 per
      cent to 1,344.75 points as of March 13, 2020. There is always more room to decline obviously due to
      the virus outbreak and oil price war," he told NST Business.
CGS-CIMB analyst Ivy Ng Lee Fang said the research firm has cut its year-end FBM KLCI target to 1,449
      points.
"We advise investors to seek shelter in defensive and high-dividend-yield stocks until the concerns over
      the global spread of COVID-19 subside.
"These, together with Malaysia's unexpected change of coalition government, could pose downside risks to
      corporate earnings, which are difficult to measure at this time," it said.
Ng said during the global financial crisis, FBM KLCI fell by 45 per cent from its peak to 829 points, its
      lowest ever and FBM KLCI earnings fell by 8.7 per cent in 2009.
She said there is a potential earning downside risk of 10.3 per cent to its current 1.6 per cent FBM KLCI
      earnings per share growth forecast if the earnings risk resembles that of during the global financial
       crisis decline (-8.7 per cent).
Meanwhile, AxiCorp market strategist Stephen Innes said small and medium enterprises (SMEs) are most at
      risk given they generally operate on small operating cushions and will need to dive into credit lines
       if not hefty interest credit card to keep things afloat.
"Unless governments actually get money physically into the hands that need it the most things can turn bad.
       This will put pressure on bank stocks due to credit line drawdown and an increase in defaults.
"We know oil and gas and airlines are in a bad way due to travel restrictions imposed around the world.
"At this stage, I see no way out and expect more investors to move into cash or keep investment portfolio
      very light primarily in recessionary proof areas like healthcare.
"But honestly with credit under duress, I think cash is the place to be. The only endgame in sight is when
       the world sees COVID-19 ends," he said.
```

Table 3 provides the explanations given by the models when inferring the values of the remaining variables during the prediction step. We can observe that the two models reason similarly, although GPT-4o tends to return a more precise and detailled explanation. The only difference in behaviour is for variable *h0* where GPT-4o does not answer. As this value is computed during the abduction step, in an anticausal fashion (predict a cause from its consequences), this is expected. This issue can be alleviated by marginalising over the set of possible values of *h0*.

Tables 4 and 5 provide examples of explanation returned by the models when evaluating the factual and counterfactual causal graphs. GPT-4o provides longer and more detailled explanations while the justification provided by LLaMA-3.1 is not always very explanative. We can see that scores and confidence are highly correlated for both models although outliers can also be observed.

Table 3: Explanation of the inferred values by GPT-4o and LLaMA-3.1 for every variable in Table 1.

	GPT-4o	LLaMA-3.1
2	The Bursa Malaysia downtrend magnitude is influenced by several factors. The severity of the COVID-19 pandemic is low, which would typically suggest a smaller downtrend magnitude. However, the severity of the oil price war is severe, which greatly increases the downtrend magnitude. The downside risks to corporate earnings are low, which would also typically suggest a smaller downtrend. Lastly, the potential end of the COVID-19 pandemic is not expected (False), which does not help reduce the downtrend magnitude. Thus, considering the severe impact of the oil price war, I predict a moderate downtrend magnitude of around 20%.	Given that Severity of oil price war is severe, increasing Bursa Malaysia downtrend magnitude and Downside risks to corporate earnings are low, contributing less to the downtrend. Although Potential end of COVID-19 pandemic would reduce the downtrend magnitude, Severity of COVID-19 pandemic being low would increase it, but this effect is expected to be lower than that of oil price war. Hence Bursa Malaysia downtrend magnitude is predicted to be high.
3	The FBM KLCI index value is influenced by several factors for the market close yesterday. The Bursa Malaysia downtrend magnitude shows a significant percentage decline from peak (20%), which has a negative impact on the index. Furthermore, the selling pressure on stocks is high, which also decreases the index value. On the other hand, the condition of the oil & gas and airlines sectors is good, which supports the index value. Given these mixed factors, the FBM KLCI index is estimated to be lower but not drastically so.	The estimated value of the FBM KLCI index is 1500.00 due to the high Bursa Malaysia downtrend magnitude and high selling pressure on stocks, which both decrease the FBM KLCI index value. However, the good condition of oil & gas and airlines sectors has a positive effect on the index value.
10	The downside risks to corporate earnings are influenced by Malaysia's change of coalition government. Since the change of coalition government did not occur (value is False), the downside risks are expected to be low.	*(no explanation returned by the model)*
11	Given that the severity of the COVID-19 pandemic is low and the potential end of the pandemic is false, there would be no severe travel restrictions imposed worldwide. The low severity of the pandemic suggests minimal disruption, and since there's no potential end of the pandemic yet, current measures would likely remain in place but not be stringent.	Given that the severity of COVID-19 pandemic is low and there is a potential end of COVID-19 pandemic, it can be inferred that travel restrictions imposed worldwide would likely be none. This is because a low severity of the pandemic suggests that the situation is under control, and a potential end to the pandemic would further reduce the need for travel restrictions.
12	Since the travel restrictions imposed worldwide were 'none' yesterday, it is likely that the oil & gas and airlines sectors condition was 'good'. Severe travel restrictions would worsen these sectors, but in the absence of any restrictions, the condition should be favorable.	Given that yesterday there were no worldwide travel restrictions, it is estimated that the condition of oil & gas and airlines sectors was good.
h0	The value cannot be determined as there are no parent causes provided to infer the outcome.	The potential end of COVID-19 pandemic will occur because all the parent causes have values that lead to this outcome.

Table 4: Example of graph scores explanations given by GPT-4o in Table 2. FG stands for Factual Graph and CG for Counterfactual Graph. Sc stands for Score and Cf stands for confidence. We show the explanations for one graph and two counterfactual graphs.

	Sc	Cf	GPT-4o
FG	0.85	0.85	The sequence of events and their causal relationships are highly plausible given the context of the COVID-19 pandemic and global economic conditions. The COVID-19 pandemic had a significant impact on global markets, including Bursa Malaysia, leading to a downtrend in FBM KLCI points. The severity of the oil price war also contributed to this downtrend. Travel restrictions due to the pandemic affected both the oil and gas sector and the airline sector, leading to further economic pressure. The change in Malaysia's coalition government affected corporate earnings, contributing to the overall economic situation. The lack of financial support from the government increased pressure on bank stocks, especially given the bad operating conditions for SMEs. The values provided for FBM KLCI points and other financial metrics are consistent with the described events and their causal relationships. Therefore, the overall sequence of events is highly plausible.
CG1	0.5	0.5	Let's break down the events and their causal relationships one by one: 1. Travel restrictions (Event 0) are false, meaning there are no travel restrictions in place. 2. The severity of the oil price war (Event 1) is high, which indicates a significant conflict impacting oil prices. 3. The severity of the COVID-19 pandemic (Event 2) is also high, suggesting widespread impact from the pandemic. 4. Travel restrictions affect the oil and gas sector condition (causal relationship 0 -> 3). 5. The condition of the oil and gas sector (Event 3) is bad. Given the provided events and relationships, the key causal link to evaluate is between travel restrictions (Event 0) and the oil and gas sector condition (Event 3). Since travel restrictions are not in place (false), this would typically suggest a neutral or positive impact on the oil and gas sector. However, the sector condition is bad. This could be due to other factors not explicitly mentioned in the graph, such as the high severity of the oil price war (Event 1) and the high severity of the COVID-19 pandemic (Event 2). The graph doesn't explicitly link these other factors to the oil and gas sector condition, but they are plausible contributors. Overall, the sequence of events seems somewhat plausible but lacks explicit causal links for all influencing factors, resulting in a moderate score and confidence.
CG2	0.3	0.5	The causal graph has several issues that impact its overall plausibility. First, the event 'FBM KLCI earnings decline in 2009' is given a value of 5.0, which is plausible given the historical context of the 2008-2009 financial crisis. However, the subsequent events 'Severity of COVID-19 pandemic' and 'Severity of oil price war' are not assigned values, making it difficult to evaluate their plausibility. The link between 'FBM KLCI earnings decline in 2009' and 'Potential earning downside risk' is plausible, as past earnings declines can influence future downside risks. However, the lack of values for the severity of the pandemic and oil price war significantly reduces the confidence in the overall plausibility of the causal graph.

Table 5: Example of graph scores explanations given by LLaMA3.1 in Table 2. FG stands for Factual Graph and CG for Counterfactual Graph. Sc stands for Score and Cf stands for confidence. We show the explanations for one graph and two counterfactual graphs.

	Sc	Cf	LLaMA-3.1
FG	0.7	0.8	The plausibility of the sequence of events is high due to the following reasons: COVID-19 pandemic (event 0) has a direct impact on recessionary trends (event 1). Fiscal policy responses (event 2) mitigated the effects of recessionary trends. However, oil price war (event 3) influenced FBM KLCI (event 4), which was also impacted by COVID-19 pandemic and recessionary trends.
CG1	0.6	0.7	The event 'COVID-19 pandemic' with a value of True increases the likelihood of recessionary trends and has an impact on FBM KLCI. The event 'Oil price war' with a value of 80 influences FBM KLCI but is not directly linked to recessionary trends, despite its potential indirect impact. However, recessionary trends are False which decreases the likelihood of FBM KLCI being influenced by it.
CG2	0.0	0.8	The COVID-19 pandemic has had a significant impact on the global economy, leading to recessionary trends. However, fiscal policy responses have mitigated some of these effects. The oil price war has also influenced the FBM KLCI, but its overall value remains high.