Exploring LLMs for Causal Discovery in Cyber-Physical Systems

Katrin Ehrenmüller[1,*,‡], Paul Maurer[1], Fajar J. Ekaputra[1], Marta Sabou[1] and Konrad Diwold[2]

[1]Vienna University of Economics and Business (WU Wien)
[2]Siemens AG Österreich

Abstract

Causality plays a fundamental role in both human reasoning and complex system analysis. As Cyber-Physical Systems (CPS) become increasingly complex, understanding causal relationships between system events is essential for tasks such as anomaly detection and fault diagnosis. This paper explores the potential of Large Language Models (LLMs) to support causal discovery in CPS. In particular, the capabilities of LLMs in assisting domain experts to identify system states and their causal relations are investigated. We propose a hybrid workflow that integrates LLM-generated suggestions with domain expert validation, aiming to improve the efficiency of causal analysis. Our evaluation is conducted on a real-world smart grid use case and compares LLM-generated causal relations with domain expert-validated ground truth. The results indicate that, while LLMs can propose relevant causal structures, their effectiveness varies depending on the complexity of temporal and topological relationships between system states. Although these models do not replace human domain expertise, they can serve as a valuable tool for supporting causal discovery in a hybrid workflow. Future research should focus on refining LLM capabilities and expanding their application across different CPS domains. Investigating different LLMs, causal models, and larger datasets may provide deeper insights into their potential for causal discovery.

Keywords

Causality, Large-Language Models, Cyber-Physical Systems, Explainability

1. Introduction

Causality has long been a topic of interest for researchers. With the rise of Machine Learning (ML) and a focus shift towards eXplainable Artificial Intelligence (XAI), the discussion of causality and how to define it has reemerged [1]. While ML models are good at explaining their choices in terms of correlations, humans use the concept of causality to explain and understand the world [2]. Causality is closely linked to our language, as we use “because”, “due to”, and similar words to express our thinking and reasoning [3].

Causality also plays an important role in the context of Cyber-Physical Systems (CPS) which integrate physical systems with computational functionalities to solve complex tasks. CPS are deployed in various industries, such as energy management [4], and manufacturing [5]. With increasing functionalities, these systems become more complex and more challenging to manage. Thus, explainability is becoming a requirement for such system designs. In the domain of CPS, explaining and finding the root causes of anomalies is crucial to detect faults in a system or to find weaknesses [6].

Causalities in terms of root cause analysis often are derived in domain expert interviews, during which domain experts try to map their natural language description of the system and their know-how to a causal model. This process is time-consuming and often difficult to create a mutual understanding of the goal of the interviews.

CausalNeSy'25: Workshop on Causal Neuro-symbolic Artificial Intelligence, June 1-2, 2025, Portoroz, Slovenia
*Corresponding author.
‡formerly published as Katrin Schreiberhuber.
katrin.schreiberhuber@wu.ac.at (K. Ehrenmüller); paul.maurer@s.wu.ac.at (P. Maurer); fajar.ekaputra@wu.ac.at (F. J. Ekaputra); marta.sabou@wu.ac.at (M. Sabou); konrad.diwold@siemens.com (K. Diwold)
http://conceptbase.sourceforge.net/mjf/ (F. J. Ekaputra); http://conceptbase.sourceforge.net/mjf/ (M. Sabou)
0000-0003-1815-8167 (K. Ehrenmüller); 0000-0001-7116-9338 (P. Maurer); 0000-0003-4569-2496 (F. J. Ekaputra); 0000-0001-9301-8418 (M. Sabou); 0000-0002-6265-4064 (K. Diwold)


As a solution approach to this problem, we use LLMs to help domain experts and causality experts, by defining a hybrid human-LLM workflow. By leveraging common knowledge capabilities as well as natural language reasoning by LLMs, domain experts can be assisted with potential causal relations in a system. Additionally, LLMs can help to translate verified knowledge to a predefined causal model, structuring domain expert knowledge into a machine-readable format. For a real-life use case in the domain of smart grids, we will test the capabilities of LLMs to discover causal knowledge in a CPS. Ground truth knowledge on causality has been collected through domain experts in multiple knowledge acquisition workshops as part of a research project. By comparing suggestions generated by LLMs with the results of domain expert workshops, we can test the capabilities of LLMs to assist in this process. Additionally, new suggestions, which have not been discovered in the domain expert workshops, will be evaluated by domain experts on their suitability in the use case.

In the remainder of the paper, we will discuss related work in Section 2, followed by a description of the use case in Section 3. A detailed explanation of the proposed hybrid causal discovery workflow is shown in Section 4. The evaluation of the approach is described in Section 5, concluding the paper in Section 6.

2. Related Work

This paper is situated in the intersection of two research areas: (1) explainable Cyber-Physical Systems, (2) LLM-augmented knowledge engineering. This section offers a brief overview of related work from each of these areas.

(Explainable) Cyber-Physical Systems. CPSs integrate computational elements with physical systems, enhancing automation, intelligence, and efficiency in various industries such as manufacturing, energy, and infrastructure [7]. However, this increased automation introduces challenges related to system transparency and explainability, making it difficult for stakeholders to understand system behavior and decision-making. Thus, the emerging research area of explainable CPS is focused on providing explanation capabilities to CPS, making these complex systems more understandable and usable to humans [6]. Initial research for ExpCPS exists for various domains, such as smart grids [8, 9, 10, 11, 12], and smart buildings [13, 14]. However, current solutions are limited to individual domains, focusing on domain-specific solutions for explainability [15]. Furthermore, explanations in CPSs rely heavily on domain knowledge from domain experts, as root cause analysis requires deep contextual understanding of a system. Reducing the time and effort that domain experts need to spend on specifying this causality knowledge is an unsolved issue, which also motivates our work. Recent work has also investigated the use of LLMs for the CPS domain. For instance, the CPS-LLM framework demonstrates how physics-aware LLMs can generate safe, context-specific plans for human-in-the-loop CPS, such as insulin dosing in diabetes management, showcasing the potential for robust, human-centric CPS control [16].

LLM-Augmented Knowledge Engineering and Causality Acquisition. LLMs have proven to be powerful tools for a wide range of tasks, including knowledge engineering and acquisition. Within knowledge engineering, LLMs have been proposed for entity extraction, link prediction, question-answering, or knowledge base construction [17]. For example, AutoKG proposes an efficient and automated approach for knowledge graph construction using LLMs. It shows, that LLMs can retrieve interconnected and comprehensive knowledge from text [18].

LLMs for Causality Acquisition More broadly, the topic of LLMs for causal discovery has gained attention recently. In [19], three key areas are highlighted: (i) *direct causal extraction from text*, (ii) *domain knowledge integration*, (iii) *causal structure refinement*. This paper is mainly situated in the domain knowledge integration. In [20], the potential of LLMs to significantly augment causal discovery is discussed, while caution should be applied to not solely trust LLM results in high-stakes decisions.

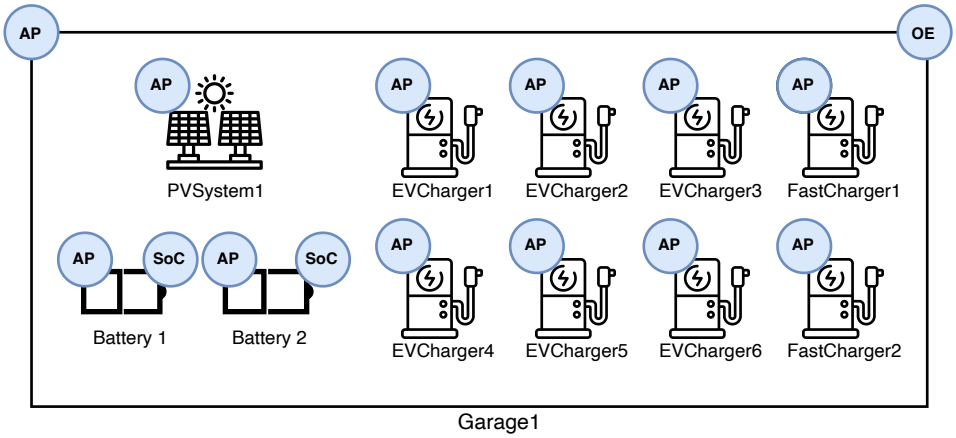

Figure 1: Schematic overview of the CPS use case - a smart charging garage with 8 EV charging stations, 2 local batteries and a PV system installed. Installed sensors are indicated as blue circles (AP = Active Power, SoC = State of Charge, OE = Operating Envelope)

The performance of LLMs mainly relies on their ability to extract common-sense information that has been present in their training data [21]. Zevcevic et al. argue that, especially for causal knowledge, LLMs cannot do inductive causal inference, as classic causal discovery models do. They argue that LLMs can reproduce causal arguments learned from their training data, while they cannot learn from actual physical measurement, as classic causal discovery methods do. Therefore, while LLMs cannot be used for causal inference on their own, they might serve as kick-starters to learning causal inference [21]. We investigate this hypothesis in this paper.

Various approaches have been proposed to augment traditional causal discovery methods via LLMs (e.g. integrating causal world models with LMs to improve performance for causal inference tasks [22], using harmonized priors to address misalignments between knowledge-based and data-implied causal relations [23], introducing causal order instead of causal graphs to address imperfect experts or LLMs [24], introducing causal modelling agents to allow a collaborative exploration of causal graph space [25, 26]). Another interesting research area is the use of small language models in combination with knowledge graph structures to improve capabilities of LMs, while saving computational efforts [27].

3. Use Case

In the evaluation of our approach, we will work on a real-life use case of an electric vehicle (EV) charging garage. The use case is represented in a knowledge graph - according to the SENSE ontology [28]. The SENSE ontology is a domain-agnostic ontology for cyber-physical systems, which covers multiple aspects of a system to enable explainability of system states. The system's setup, topology and existing sensors are represented in a well-defined, reusable and publicly accessible[1] data model, which can be reused for other use cases as well. Thus, our experiments can be replicated for any other use cases, where the system is represented according to the SENSE ontology.

In the SENSE ontology, causal relations are defined, using three layers: (i) type of causality, (ii) temporal relations, (iii) topological relation. These three aspects help in defining causal relations in the context of CPS, as they are very much dependent on temporal and topological proximity of two states within a system. These intricacies of a causal relation are crucial to derive accurate root causes of a state in a running system [28].

(i) Causality in CPS is commonly defined by domain experts. Thus, causality is largely defined through natural language. Based on the CaTeRS schema [3], we have identified three types of

[1] http://w3id.org/explainability/sense#

causality from natural language, which are relevant for describing causality in the context of CPS: *cause* ($A \rightarrow B$), *enable* ($\neg A \rightarrow \neg B$), *prevent* ($A \rightarrow \neg B$).

(ii) The temporal aspect of a causal relation defines the temporal restrictions of a causal relation to hold. This restriction can define, whether two states need to happen at the same time (*overlap*, or *identity*), or if they can be temporally distinct (*before*). This definition also comes from the CaTeRS schema [3], defining causal relations in text.

(iii) The topological aspect defines restrictions on the proximity of two states for a causal relation to hold. Thus, it defines, which platform can be affected be the cause state (*samePlatform*, *siblingPlatform*, or *parentPlatform*).

The use case is comprised of a public garage, which offers charging stations for electric vehicles. As the garage is a major energy consumer in the power grid, an operating envelope is imposed on the garage by the local power grid provider. An operating envelope defines a power limit, which can be used at a given point in time. If this limit is violated by the garage, operating envelope violation event occurs. Such a violation constitutes a problem for the facility operator of the garage (the facility operator violates their contract with the power grid provider), as well as the power grid provider (they need to deal with an unexpected peak in power consumption, which can be dangerous to the local energy grid, if it is not mitigated). Therefore, potential causes of an operating envelope violation are vital explanations to reduce fees for the facility operator of the garage, and ensure the stability of the power grid.

The smart charging garage is equipped with a set of sensors, to measure the power consumption of different devices. In Figure 1, the use case setup with devices and their sensors is shown. The garage contains 8 EV charging stations and 2 batteries, which can be used for peak shaving (reducing the power consumption from the local grid by providing power from the battery). Additionally, there is a photovoltaic (PV) system at the roof of the garage, which can be used to recharge the battery, to directly charge an EV (if one is connected to a charger), or to feed back to the local power grid. At the garage level, there is a sensor, which sends the current limit of the imposed operating envelope.

For this use case, a list of states as potential causes for an operating envelope violation should be collected. This list of potentially interesting states can then be used further to analyse which of the potential causes has happened close to the violation - finding the actual cause of of a specific violation at a time point *t*. In case of an envelope violation, a helpful explanation for the facility operator could look like this:

> "*EnvelopeViolationX* at *Garage1* happened at *t*. This violation was caused by *HighChargingY* at *FastCharger1*, which started at *t* and was enabled by *LowBatteryStateOfChargeZ* at *Battery1* at *t*."

For such an explanation, causal relations from experts are needed. In this example explanation specifically, the following causal relations are used, as defined in the SENSE ontology:

> "*HighCharging* causes overlapping *EnvelopeViolation* at parentPlatform."
> "*LowBatterySoC* enables overlapping *EnvelopeViolation* at parentPlatform."

In this paper, the acquisition of these causal relations is investigated, looking into how LLMs can assist experts in this causal discovery process.

4. LLM-Enhanced Causal Discovery Workflow

We propose an LLM-enhanced causal discovery approach to find causal relations between CPSs system states. An LLM prompting workflow is defined to help domain experts by suggesting causal relations in their CPS to kickstart the work on causal discovery by domain experts. In Figure 2, a traditional

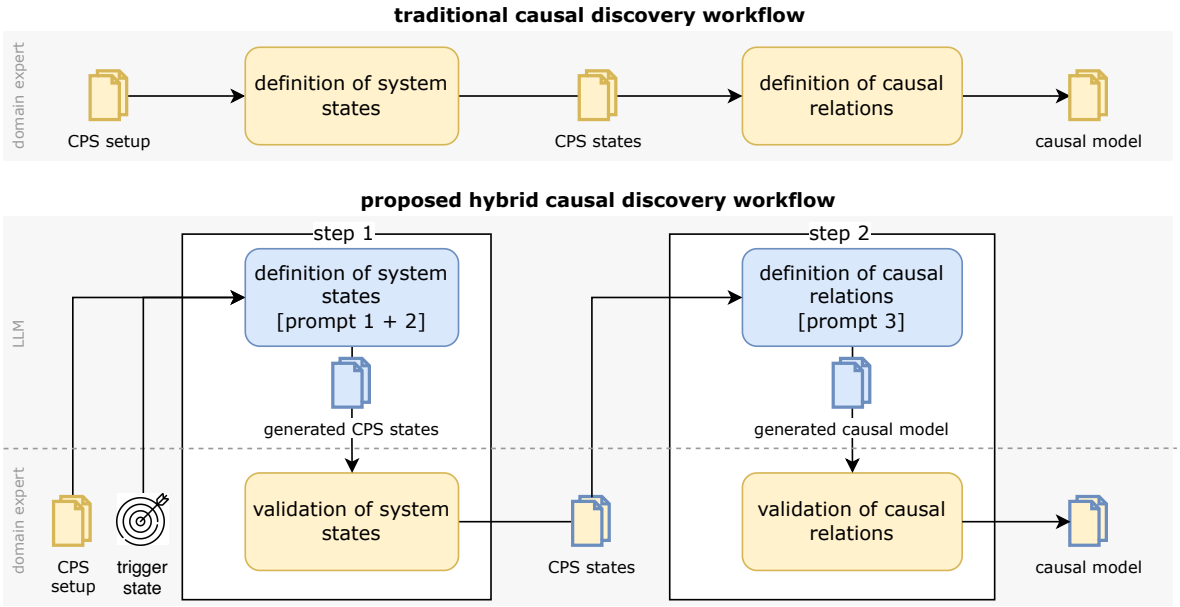

Figure 2: Traditional vs proposed causal discovery workflow to define causal relations in CPS. Yellow tasks represent domain expert responsibilities, blue tasks represent LLM responsibilities.

causal discovery workflow is depicted at the top, where all tasks are in the responsibility of the domain expert: defining the possible states that can happen in a system, as well as the definition of causal relations between these states. We propose a workflow (Figure 2 bottom), where LLMs are leveraged to define potential system states and relations. In this workflow, a domain expert is merely responsible for validating the output of the LLM. This approach can help in the efficiency of the process, as domain experts are presented with an initial idea on the type of relations that could be relevant in their system.

The workflow is developed based on prompt patterns defined by White et al [29]. The main sections of each prompt are shown in this paper (see Prompt 1-3). For further clarifications and the full prompts, please refer to our github repository[2].

The proposed workflow is split into two major steps: *Step 1* focuses on the definition of potential system states, which can occur in a CPS. As an example, in the EV charging garage use case, potential states would be *LowStateOfCharge* of a battery, or an *EnvelopeViolation*. *Step 2* aims to define causal relations between the states that have been defined in Step 1. In our use case example, this would be *"LowStateOfCharge enables EnvelopeViolation at parent platform, if these states overlap"*. As it is shown in the example, a detailed definition of causal relations is employed in this use case. CPSs rely mostly on sensor data from sensors deployed across the full system. Therefore, additional parameters and constraints need to be added to a causal relation to define accurate causalities in a system. A more detailed description of these added constraints is provided in step 2.

Step 1 - state definition. In this initial step, the LLM is presented with all relevant information to understand the system. Some general terms are defined, and system setup data is given as input to the LLM. Additionally, the domain expert is required to define a trigger state - a state they are interested in. This means, they need to define the state they want to explain in their system (e.g., in our use case the envelope violation). A trigger state defines the area of interest for defining and detecting states, within a CPS. Then, the LLM is asked for a set of states, which could be related to the trigger state. This step is defined by two prompts in a chain-of-thought experiment [30]. In Prompt 1, the main sections of the data input prompt are shown. The goal is to make sure that the system input is well understood by the LLM. At this stage, misunderstandings can be rectified by the domain expert, if needed. In Prompt 2,

[2]https://anonymous.4open.science/r/causaldiscoverywithLLMs-5EAF/

the system setup should be used to create possible state types, which could be interesting to a user to identify the root cause of the given trigger state.

Prompt 1: Data input and understanding

```
Let's define a concise ``meta-language" for clarity.
<define relevant terms and relations>
I have tabular data, and I would like it transformed into a descriptive natural
 language format where each row is expressed as a sentence.
<...>
Use your general knowledge and reasoning capabilities to decide the relation
between the columns. It does not necessarily have to be ``is a type of", but
could also reflect a totally different relation depending on the context of the
 data.
<...>
Please summarize the provided data and locate it within the context of related
topics. If there are issues, let me know.
```

Prompt 2: Definition of system states

```
Using the structured data from the first prompt, generate StateTypes
that describe system conditions based on observable properties and platform
types. Ensure that each StateType reflects an actual system state that can be
used for causal inference.
We are mainly interested in states connected to the cause of:
<define and describe trigger state>
<...>
Now, generate additional relevant StateTypes. Ensure accuracy in
platform association and describe each state in detail. The output format has
to be a table.
```

Step 2 - causal discovery. Building on a set of defined states from step 1, in step 2 the LLM is prompted to find causal relations between the proposed states, which could be relevant in the system. We base our causal relations on the definition of causal relations in CPS from the SENSE ontology.

Other causal models can be used in future experiments to investigate whether LLM performance changes. Prompt 3 shows the design of the second step, where the LLM is asked for causal relations between the states that have been proposed and defined in the previous step. The causal model is defined as part of Prompt 1, where relevant terms and relations are defined as a "meta-language" for the workflow.

Prompt 3: definition of causal relations

```
Using the predefined StateTypes, generate causal dependencies that
describe how system states influence one another. Each causal dependency should
 follow this format:
StateType_cause, Causal Relation, Temporal Relation, Topological Relation,
StateType_effect
StateTypeA, causes, overlaps, parentPlatform, StateTypeB

Now, generate relevant causal dependencies based on the following StateTypes:
<define state types to be used>
Both the Cause State and the Effect State must be strictly selected from the
predefined StateTypes. No new StateTypes should be created in this step. Only
causal dependencies between existing StateTypes should be established.
Ensure accuracy in temporal and topological associations. If any concept is
unclear, highlight uncertainties and request expert clarification to ensure
correct understanding.
```

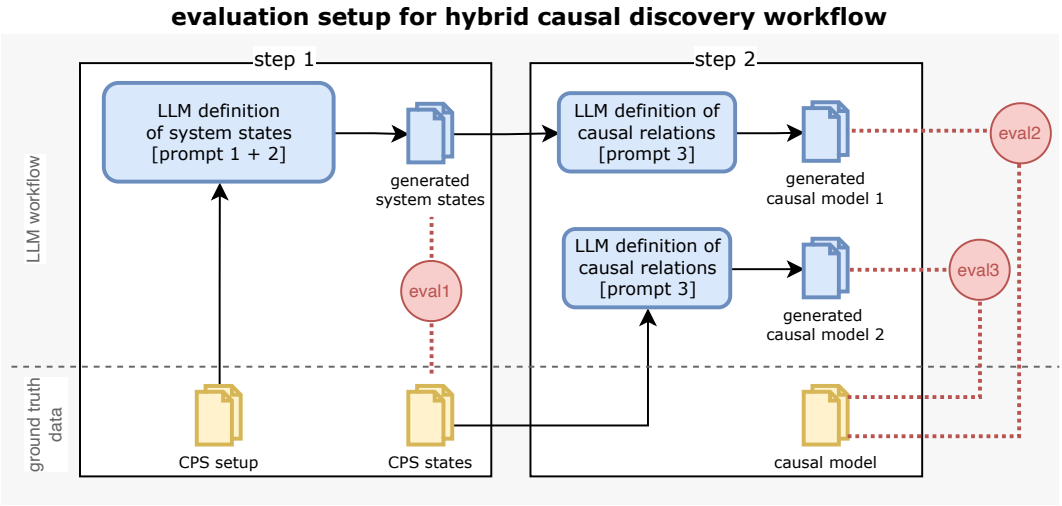

Figure 3: Evaluation setup of the hybrid causal discovery workflow (yellow = ground truth data from domain experts, blue = LLM-related artefacts, red = evaluations).

5. Evaluation

The workflow proposed in Section 4 is evaluated against ground truth data from domain experts, which has been collected in a research project through multiple workshops over the course of a year. We test the performance of GPT-4o [31] (using the default temperature setting of 0.7) and deepseek-R1 [32] (using the default temperature setting of 1.0). For one use case, the workflow is tested in two settings: (i) run the full workflow with the LLM, without interference (evaluation 1+2). (ii) test the steps individually, using ground truth as input data (evaluation 1+3).

The correctness of each proposed state, or causal relation, of an LLM is labeled manually by a domain expert. Since the naming of states is different for each LLM and domain experts, manual matching based on the state description has to be conducted, to facilitate the labeling process. There are three possible labels for a proposed instance.

- *true*: equivalent to an instance that is part of the ground truth.
- *reasonable*: not part of the ground truth, but the instance is a reasonable suggestion according to domain experts.
- *false*: not a reasonable suggestion.

Based on these labels, we measure precision and recall on ground-truth data of the use case. In the scope of this paper, precision considers the sum of all true and reasonable instances proposed by an LLM, divided by the total number of suggestions by an LLM. Thus, the percentage of useful proposals is calculated. Recall is calculated by dividing the number of true suggestions by the number of ground truth elements. This measure shows the overlap between domain expert suggestions and LLM suggestions.

$$precision = \frac{true + reasonable}{totalSuggestions} \qquad recall = \frac{true}{totalGroundtruth}$$

As we do not assume our ground truth to exhaustively include all possible states and causal relations, these metrics allow us to measure the performance of LLMs, even beyond the ground truth data that has been collected from domain experts.

Definition of system states. First, the generation of system states from a system setup is tested with both LLMs (GPT-4o and deepseek-R1), which is the result of step 1 in the workflow (prompts 1 and 2). In Table 1, precision and recall for the generation of system states is shown for GPT-4o and deepseek-R1. These are the performance metrics of evaluation 1. While recall is relatively low for both models (0.55

	evaluation 1	
	precision	recall
GPT-4o	0.91	0.55
deepseek-R1	0.69	0.45

Table 1
Performance of step 1 in the workflow (definition of system states).

		evaluation 2		evaluation 3	
		precision	recall	precision	recall
(1) cause-effect pair	GPT-4o	0.80	0.20	0.83	0.30
	deepseek-R1	0.77	0.00	0.67	0.10
(2) causal relation	GPT-4o	0.60	0.10	0.75	0.20
	deepseek-R1	0.77	0.00	0.50	0.20
(3) temporal relation	GPT-4o	0.70	0.10	0.50	0.00
	deepseek-R1	0.54	0.00	0.67	0.20
(4) topological relation	GPT-4o	0.20	0.20	0.64	0.10
	deepseek-R1	0.62	0.00	0.67	0.20

Table 2
Performance of step 2 in the workflow (definition of causal relations).

for GPT, and 0.45 for deepseek), precision of both models shows promising results (0.91 for GPT and 0.69 for deepseek). In this evaluation, it can be shown that replicating the manual work of domain experts cannot be achieved in a one-shot workflow. However, new and potentially interesting states can be proposed, which might help in discovering states, which have not been considered by domain experts yet. Creating a set of system states, where more than two thirds of the proposed options are suitable for further investigation can help domain experts, especially in a first step to consider potential states to investigate.

Definition of causal relations. This step was evaluated using two different setups: *Evaluation 2* was conducted by using the list of generated states from step 1. *Evaluation 3*, instead, was conducted by using the ground truth states, defined by domain experts, as the input data. Each evaluation was tested with two LLMs, calculating precision and recall, as previously defined. As defined in Section 2, four different aspects of causality were captured in this step. We evaluated each aspect separately, to find strengths and weaknesses of LLMs for each relation, as follows:

- *(1) cause-effect pair* captured the correctness of a relation between two states in general (i.e. stateA has some causal influence on stateB).
- *(2) causal relation* measures the correctness of the type of causality proposed by a model (i.e, one of cause, enable, prevent).
- *(3) temporal relation* is concerned with the temporal correctness of the causality (overlaps, before).
- *(4) topological relation* considers whether the spatial relation proposed by an LLM is correct for the causal relation (samePlatform, siblingPlatform, parentPlatform).

Naturally, performance decreases from (1) to (4), as correct cause-effect pairs are prerequisites for correct types of relations.

Using generated states for creating a causal model (evaluation 2) showed promising results in terms of precision, when looking at cause-effect pairs (0.80 for GPT, 0.77 for deepseek). Interestingly, this setup also outperformed using ground truth data as input for defining causal relations. However, for temporal relations and topological relations, the comparison is not so clear. For recall, performance is naturally better, when using ground truth states (with the exception of temporal relations and topological relations for GPT). When using ground truth data, there are more relevant states, which can be used by the LLMs to define causal relations. Generally, the results suggest that proposing temporal and topological relations is considerably harder for GPT than proposing cause-effect pairs and causal relations. While GPT performs best for category (1), decreasing continuously for (2), (3) and (4), deepseek is more consistent in its performance over all categories (in terms of precision).

Overall, these results do not give a conclusive answer to which LLM performs best for the task of causal discovery. Additionally, current results are not sufficient to replace human domain experts in

the process of causal discovery. However, LLMs can create instances of causal connections, which are correct (or at least reasonable) at least 67% of the time - with decreasing performances, when defining more details on a causal relation (defining causal, temporal and topological relation). This shows that, while LLMs cannot fully replace humans, they are capable of proposing a set of relations, which can be used to support the causal discovery process performed by (domain) experts. Yet, the most impressive finding in this setup is the fact that all of these causal relations are defined by LLMs from very limited data, merely by knowing about the system setup and one trigger state, which should be explained. There is no input on how variables influence each other, or how certain variables are measured/calculated. In comparison, starting from such a small set of data requires domain experts to discuss possible scenarios for multiple hours, before they can define causal relations (as was done during the process of creating the ground truth dataset).

6. Conclusion

Causality has been a key concept in both human reasoning and complex system analysis. With the increasing complexity of Cyber-Physical Systems (CPS) and their applications, such as smart grids, causal relationships have become even more relevant, especially for applications, such as root cause analysis. This paper explored the potential of leveraging Large Language Models (LLMs) to assist domain experts in discovering causal relations in such systems.

Our proposed hybrid causal discovery workflow, where LLMs generate potential system states and causal relations, offers a promising approach to enhance the efficiency of causal discovery in CPS. By validating the suggestions from LLMs against domain expert knowledge, we found that these models can provide valuable initial suggestions, even though they are not capable of replacing human domain expertise at the moment. Evaluation results demonstrated that LLMs such as GPT-4o and deepseek-R1 can successfully suggest system states and causal relations, with precision rates showing potential for further use in domain expert-led causal discovery processes.

Although LLMs performed well in identifying cause-effect pairs and some aspects of causal relations, challenges remain in capturing temporal and topological nuances of causality. This indicates the need for further refinement, particularly in the context of CPS, where these temporal and spatial considerations are essential for accurate analysis.

The hybrid workflow presented in this study provides a solid foundation for integrating machine learning into the causal discovery process. Future research could explore ways to enhance an LLM's understanding of complex temporal and topological relationships, as well as investigate their potential applications in broader CPS contexts. Additionally, further experimentation with different LLMs, causal models and larger datasets will help in understanding the limits and potential of LLMs in causal analysis in the context of CPS. Investigating a more integrated workflow between LLMs and domain experts could also constitute a valuable contribution towards a more efficient and ML-assisted causal discovery process. Furthermore, testing this workflow in less-known domains (e.g. aerospace engineering, or specific manufacturing processes) could be used to show an LLM's ability to reason over knowledge and data it has not been trained on.

In conclusion, while LLMs cannot replace domain experts, they can assist in the early stages of causal discovery, providing valuable insights that can guide domain expert-driven analysis and decision-making in the context of complex CPS.

References

[1] G. Carloni, A. Berti, S. Colantonio, The role of causality in explainable artificial intelligence, arXiv preprint arXiv:2309.09901 (2023).

[2] S.-H. Lin, M. A. Ikram, On the relationship of machine learning with causal inference, European journal of epidemiology 35 (2020) 183–185.

[3] N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, L. Vanderwende, CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures, in: Proceedings of the Fourth Workshop on Events, Association for Computational Linguistics, San Diego, California, 2016, pp. 51–61. URL: https://aclanthology.org/W16-1007. doi:10.18653/v1/W16-1007.
[4] X. Yu, Y. Xue, Smart grids: A cyber–physical systems perspective, Proceedings of the IEEE 104 (2016) 1058–1070.
[5] L. Monostori, B. Kádár, T. Bauernhansl, S. Kondoh, S. Kumara, G. Reinhart, O. Sauer, G. Schuh, W. Sihn, K. Ueda, Cyber-physical systems in manufacturing, Cirp Annals 65 (2016) 621–641.
[6] S. S. Jha, An Overview on the Explainability of Cyber-Physical Systems, The International FLAIRS Conference Proceedings 35 (2022). URL: https://journals.flvc.org/FLAIRS/article/view/130646. doi:10.32473/flairs.v35i.130646.
[7] H. A. Müller, The rise of intelligent cyber-physical systems, Computer 50 (2017) 7–9.
[8] J. E. Larsson, B. Öhman, A. Calzada, Real-time root cause analysis for power grids, in: Security and Reliability of Electric Power Systems, CIGRE Regional Meeting, Tallinn, Estonia, Citeseer, 2007.
[9] P. R. Aryan, F. J. Ekaputra, M. Sabou, D. Hauer, R. Mosshammer, A. Einfalt, T. Miksa, A. Rauber, Simulation support for explainable cyber-physical energy systems, in: 2020 8th Workshop on Modeling and Simulation of Cyber-Physical Energy Systems, IEEE, 2020, pp. 1–6.
[10] P. R. Aryan, F. J. Ekaputra, M. Sabou, D. Hauer, R. Mosshammer, A. Einfalt, T. Miksa, A. Rauber, Explainable cyber-physical energy systems based on knowledge graph, in: Proceedings of the 9th Workshop on Modeling and Simulation of Cyber-Physical Energy Systems, 2021, pp. 1–6.
[11] J. Cordova, L. M. K. Sriram, A. Kocatepe, Y. Zhou, E. E. Ozguven, R. Arghandeh, Combined electricity and traffic short-term load forecasting using bundled causality engine, IEEE Transactions on Intelligent Transportation Systems 20 (2018) 3448–3458.
[12] A. Chhokra, N. Mahadevan, A. Dubey, G. Karsai, Qualitative fault modeling in safety critical cyber physical systems, in: Proceedings of the 12th System Analysis and Modelling Conference, 2020, pp. 128–137.
[13] C. Lork, V. Choudhary, N. U. Hassan, W. Tushar, C. Yuen, B. K. K. Ng, X. Wang, X. Liu, An ontology-based framework for building energy management with iot, Electronics 8 (2019) 485.
[14] J. Ploennigs, A. Schumann, F. Lécué, Adapting semantic sensor networks for smart building diagnosis, in: The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13, Springer, 2014, pp. 308–323.
[15] J. Greenyer, M. Lochau, T. Vogel, Explainable software for cyber-physical systems (es4cps), in: GI Dagstuhl Seminar, volume 19023, 2019.
[16] A. Banerjee, A. Maity, P. Kamboj, S. K. S. Gupta, CPS-LLM: Large Language Model based Safe Usage Plan Generator for Human-in-the-Loop Human-in-the-Plant Cyber-Physical System, 2024. URL: http://arxiv.org/abs/2405.11458. doi:10.48550/arXiv.2405.11458, arXiv:2405.11458 [cs].
[17] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities, World Wide Web 27 (2024) 58.
[18] B. Chen, A. L. Bertozzi, Autokg: Efficient automated knowledge graph generation for language models, in: 2023 IEEE International Conference on Big Data (BigData), IEEE, 2023, pp. 3117–3126.
[19] G. Wan, Y. Lu, Y. Wu, M. Hu, S. Li, Large Language Models for Causal Discovery: Current Landscape and Future Directions, 2025. URL: http://arxiv.org/abs/2402.11068. doi:10.48550/arXiv.2402.11068, arXiv:2402.11068 [cs].
[20] E. Kıcıman, R. O. Ness, A. Sharma, C. Tan, Causal Reasoning and Large Language Models: Opening a New Frontier for Causality (????).
[21] M. Zečević, M. Willig, D. S. Dhami, K. Kersting, Causal parrots: Large language models may talk causality but are not causal, arXiv preprint arXiv:2308.13067 (2023).
[22] J. Gkountouras, M. Lindemann, P. Lippe, E. Gavves, I. Titov, Language agents meet causality–bridging llms and causal world models, in: accepted as ICLR 2025 Poster, 2025.
[23] T. Ban, L. Chen, D. Lyu, X. Wang, Q. Zhu, H. Chen, LLM-Driven Causal Discovery via Harmonized

Prior, IEEE Transactions on Knowledge and Data Engineering 37 (2025) 1943–1960. URL: https://www.computer.org/csdl/journal/tk/2025/04/10839116/23tgS0cIwvu. doi:10.1109/TKDE.2025.3528461, publisher: IEEE Computer Society.

[24] A. Vashishtha, A. G. Reddy, A. Kumar, S. Bachu, V. N. Balasubramanian, A. Sharma, CAUSAL ORDER: THE KEY TO LEVERAGING IMPERFECT EXPERTS IN CAUSAL INFERENCE (2025).

[25] A. Abdulaal, A. Hadjivasiliou, A. Ijishakin, I. Drobnjak, D. C. Castro, D. C. Alexander, CAUSAL MODELLING AGENTS: CAUSAL GRAPH DIS- COVERY THROUGH SYNERGISING METADATA- AND DATA-DRIVEN REASONING (2024).

[26] H. D. Le, X. Xia, Z. Chen, Multi-Agent Causal Discovery Using Large Language Models, 2024. URL: http://arxiv.org/abs/2407.15073. doi:10.48550/arXiv.2407.15073, arXiv:2407.15073 [cs] version: 1.

[27] Y. Susanti, M. Färber, Knowledge Graph Structure as Prompt: Improving Small Language Models Capabilities for Knowledge-based Causal Discovery, 2024. URL: http://arxiv.org/abs/2407.18752. doi:10.48550/arXiv.2407.18752, arXiv:2407.18752 [cs].

[28] K. Schreiberhuber, F. J. Ekaputra, M. Sabou, D. Hauer, K. Diwold, T. Frühwirth, G. Steindl, T. Schwarzinger, Towards a state explanation framework in cyber-physical systems, in: DACH+ Energy Informatics 2024, Lugano, Switzerland, Proceedings, ACM SIG Energy, 2024.

[29] J. White, S. Hays, Q. Fu, J. Spencer-Smith, D. C. Schmidt, Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design, in: Generative AI for Effective Software Development, Springer, 2024, pp. 71–108.

[30] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[31] O. et al., Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[32] D.-A. et al., Deepseek-v3 technical report, 2024. URL: https://arxiv.org/abs/2412.19437. arXiv:2412.19437.