Early Insights into Argumentation-Guided Causal Evaluation with the Help of LLMs

Pietro Baroni[1], Federico Cerutti[1,2,3,*], Massimiliano Giacomin[1], Gian Franco Lamperti[1] and Marina Zanella[1]

[1]DII - Universitá di Brescia - Italy
[2]Cardiff University - UK
[3]University of Southampton - UK

Abstract

The rapid growth of Deep Neural Networks (DNNs) has brought substantial advances in artificial intelligence across domains such as vision, language, and recommendation systems. However, this progress comes at a steep energy cost, with model training and deployment contributing significantly to global computational energy consumption. Understanding what drives this energy demand requires more than empirical correlation—it demands causal explanations. In this work, we investigate the causal factors underlying energy use in DNN training, using structure learning algorithms such as the PC algorithm to derive candidate causal graphs. Recognising the limitations of such methods—particularly in terms of assumptions and finite data—we introduce a novel approach to evaluate each inferred link through formal argumentation. We treat each proposed causal relationship as a dialectical object, generating arguments and counterarguments that articulate its plausibility, underlying mechanisms, and possible confounders. We operationalise this reasoning using large language models in a zero-shot prompting setup, surfacing the evidential and conceptual assumptions behind each causal claim. This hybrid approach, combining causal discovery with structured argumentative evaluation, promotes interpretability and critical scrutiny in data-driven causal modelling. Preliminary results demonstrate its potential for rendering causal claims more transparent and contestable.

Keywords

LLM, Causality, Argumentation

CEUR-WS.org/Vol-4025/paper1.pdf

1. Introduction

Deep Neural Networks (DNNs) have become foundational to the current landscape of artificial intelligence (AI), enabling advances in fields as diverse as computer vision [1], natural language processing [2], personalised recommendation [3], and speech recognition [4]. These models are typically trained on large datasets using high-performance GPU clusters, often within large-scale data centres. As a consequence, the growth of deep learning research and applications has been accompanied by a substantial increase in the energy required to train and operate these models [5].

Recent analyses show that this trend, dubbed the "Red AI" era [6, 7, 8], presents a mounting ecological challenge. AI models are growing in size, complexity, and resource demand at an exponential rate. Despite ongoing improvements in chip efficiency and data centre cooling, the energy consumption and carbon emissions associated with training large models continue to double every 4–6 months. A projection [9] suggests that computing could account for up to 20.9% of global electricity demand by 2030. This includes contributions from both training and inference workloads.

To better understand the drivers of such energy consumption, we need to move beyond simple empirical correlations and ask causal questions: What parameters truly influence energy use during model training? Does batch size cause higher energy demand, or is it a proxy for another latent factor? Addressing such questions calls for causal discovery—the process of inferring causal relationships from data. Algorithms such as the PC algorithm (Section 2) and its variants provide a data-driven

AI^3 *2025: 9th Workshop on Advances in Argumentation in Artificial Intelligence, September 13, 2025, Rende, Italy*
*Corresponding author.
federico.cerutti@unibs.it (F. Cerutti)
0000-0003-0755-0358 (F. Cerutti)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

means to construct causal graphs, identifying conditional independencies and orienting plausible causal directions.

Yet, as with any form of inference, these algorithmic outputs raise as many questions as they answer. Given the statistical limitations of finite data and the assumptions encoded in the discovery procedures, the resulting causal links should not be accepted uncritically. Instead, each proposed causal relationship demands a careful evaluation of its plausibility and justification—not only in light of data, but also in terms of domain knowledge, mechanisms, and conceptual coherence.

This paper proposes that tools from formal argumentation theory [10, 11] can help operationalise such evaluations (Section 3). In our setting, each candidate causal link is treated as an object of dialectical discussion: we construct arguments in favour of the link (*e.g.*, based on observed regularities, known physical constraints, or plausible mechanisms) and contrast them with counterarguments (*e.g.*, suggesting alternative explanations, questioning generalisability, or highlighting confounding factors).

We implement this deliberative process (Section 4) using large language models (LLMs) in a zero-shot prompting setting [12]. For each candidate edge in a data-derived causal graph, the system generates structured natural language arguments and counterarguments. Rather than discarding links or altering the graph, we aim to surface the implicit reasoning behind each relation—bringing to light the assumptions, analogies, evidence patterns, and objections that might otherwise remain hidden.

The contribution of this work is a method for making the structure and justification of causal claims explicit, interpretable, and open to contestation. In doing so, we offer a hybrid approach to causal inference that supplements algorithmic discovery with argumentative reasoning. We discuss our preliminary results in Section 5 and address the limitations of the methodology in Section 6.

2. Causal Discovery

Causal discovery aims to infer causal relationships from data, typically under structural assumptions. A foundational framework for causal reasoning is provided by Judea Pearl's theory [13, 14, 15], formalised using directed graphical models and structural equations.

2.1. Structural Causal Models

The formal model used in Pearl's framework is typically referred to as a *Structural Causal Model* (SCM) or a *Directed Graphical Causal Model* (DGCM). It combines a directed acyclic graph (DAG) with a joint probability distribution that encodes assumptions about direct causal mechanisms and the independence structure of the system.

Definition 2.1 (Structural Causal Model)**.** A *Structural Causal Model* (SCM) is a tuple $\mathcal{M} = \langle \mathcal{G}, \mathcal{X}, \mathcal{F}, P(U) \rangle$ where:

- $\mathcal{G} = (V, E)$ is a directed acyclic graph (DAG), where each node $X_i \in V$ represents a variable.
- $\mathcal{X}$ is the set of possible values for the variables.
- $\mathcal{F} = \{f_i : \mathrm{Pa}(X_i) \times U_i \to X_i\}$ is a set of functions, one for each variable X_i, where $\mathrm{Pa}(X_i)$ denotes the set of parents of X_i in $\mathcal{G}$ and U_i is an exogenous noise variable.
- $P(U) = \prod_i P(U_i)$ is a product distribution over the exogenous variables.

Each endogenous variable X_i is determined by $X_i := f_i(\mathrm{Pa}(X_i), U_i)$.

Definition 2.2 (Intervention and do-calculus)**.** Given a SCM $\mathcal{M}$ and a variable $X_i \in V$, the effect of an intervention $\mathrm{do}(X_i = x)$ is a modified model $\mathcal{M}_{\mathrm{do}(X_i=x)}$ where the structural function f_i is replaced by the constant function $f_i := x$, and the distribution over other variables is adjusted accordingly.

Definition 2.3 (Causal Bayesian Network)**.** A *Causal Bayesian Network* is a pair $(\mathcal{G}, P)$ where:

- $\mathcal{G}$ is a DAG.
- P is a joint probability distribution over the variables such that the *Markov condition* holds: each variable is independent of its non-descendants given its parents in the graph.

2.2. D-separation and Faithfulness

To read conditional independencies from a graph, we rely on the notion of d-separation.

Definition 2.4 (D-separation)**.** Let $\mathcal{G}$ be a DAG and let X, Y, and Z be disjoint sets of nodes in $\mathcal{G}$. A path between a node in X and a node in Y is said to be *blocked* by Z if any of the following holds:

- The path contains a *non-collider* node that is in Z. A node v on a path is a non-collider if the path traverses it via either $u \to v \to w$, $u \leftarrow v \leftarrow w$, or $u \leftarrow v \to w$.
- The path contains a *collider* node v (i.e., a node where the arrows on the path converge: $u \to v \leftarrow w$), and neither v nor any of its descendants are in Z.

We say that X is *d-separated* from Y given Z if all paths between any node in X and any node in Y are blocked by Z.

Definition 2.5 (Causal Markov Assumption)**.** A distribution P over a set of variables V satisfies the *Causal Markov Assumption* with respect to a DAG $\mathcal{G}$ if every variable is conditionally independent of its non-descendants given its parents.

Definition 2.6 (Causal Faithfulness Assumption)**.** A distribution P is *faithful* to a DAG $\mathcal{G}$ if every conditional independence relation that holds in P is entailed by d-separation in $\mathcal{G}$.

2.3. Constraint-Based Causal Discovery and the PC Algorithm

Under the Causal Markov and Faithfulness assumptions, we can discover aspects of the causal structure by testing for conditional independencies in the observed data. One of the most widely used methods in this class is the *PC algorithm* [16].

The PC algorithm is *sound* under the Causal Markov and Faithfulness assumptions, and *pointwise consistent* under large sample limits and no latent confounding variables. It outputs a *Completed Partially Directed Acyclic Graph* (CPDAG), which represents the Markov Equivalence Class (MEC) of all DAGs consistent with the observed independencies.

Despite its widespread adoption in causal discovery research, the PC algorithm exhibits several important limitations that must be acknowledged. First, it operates under the assumption that there are no latent confounding variables or selection biases in the data. This means that the algorithm presumes all relevant variables influencing the observed relationships have been measured and are included in the analysis. In real-world applications, especially in domains like medicine, economics, or machine learning system diagnostics, such assumptions are often violated. Hidden confounders may induce spurious dependencies or mask true causal links, thereby compromising the validity of the inferred causal structure.

Second, the PC algorithm relies critically on the correctness of statistical tests for conditional independence. These tests serve as the foundation for removing edges and orienting v-structures during the graph construction process. However, in finite samples, such tests are prone to both Type I and Type II errors. A *Type I error* (false positive) occurs when the test incorrectly rejects the null hypothesis of conditional independence when it is in fact true, while a *Type II error* (false negative) occurs when the test fails to reject the null hypothesis despite the variables being dependent. These risks are exacerbated when conditioning on large sets of variables, which increases variance, or when the dependencies are weak or non-linear, making them harder to detect. As a consequence, the structure returned by the PC algorithm may be unstable or incorrect if the statistical decisions deviate from the population-level conditional independencies.

Third, the algorithm cannot distinguish between different directed acyclic graphs (DAGs) that belong to the same Markov Equivalence Class (MEC). By design, it returns a Completed Partially Directed Acyclic Graph (CPDAG), which encodes a set of DAGs that share the same set of d-separation relations. While this is theoretically sound, it limits the informativeness of the output: many edges may remain undirected, and causal directionality is left ambiguous without additional assumptions or interventional data.

Finally, and perhaps most crucially from the perspective of formal argumentation, the PC algorithm is entirely syntactic in its use of statistical dependencies; it provides no account of the semantic content of the edges in the resulting graph. Each link is treated purely as an artefact of probabilistic association (or lack thereof) rather than as a meaningful causal hypothesis embedded in domain knowledge. This poses a substantial challenge in applications where interpreting, justifying, or contesting individual links is essential. Therefore, beyond purely algorithmic discovery, there is a pressing need for frameworks that allow for the articulation and critical evaluation of causal claims—link by link—using structured arguments, explanations, and counter-arguments. Such argumentative approaches can complement statistical discovery methods by making the assumptions, interpretations, and domain-specific implications of each inferred causal relationship explicit and debatable.

3. A Structured Account of Arguments For and Against Causal Claims

Causal arguments are central in many domains, from scientific inquiry to policy and engineering, where actors seek to justify why one event (the effect) follows from another (the cause). This section provides a structured typology of argumentative forms both in support of and against causal claims, building on top of [10, 11].

3.1. Arguments in Favour of Causal Claims

Arguments supporting causal links can be grouped by the type of justification they provide. We distinguish three principal classes: circumstantial evidence, contrastive evidence, and causal explanations.

3.1.1. Circumstantial Evidence

Circumstantial evidence relies on regularities or proximity relations between two events, without offering a mechanistic explanation of counterfactual analysis. The reasoning typically takes the form:

> "A caused B because A regularly precedes B, co-occurs with B, or resembles other known causes of B."

We distinguish several sub-types of circumstantial evidence:

Spatio-temporal contiguity: The cause and effect occur in close spatial or temporal proximity. This supports the intuition that the proximity of events may suggest a causal link.

Repeated co-occurrence: The purported cause and effect consistently appear together across multiple instances. This statistical regularity, though not sufficient for causation, can signal a potentially robust association worth investigating.

Analogical similarity: The situation under analysis resembles other known causal scenarios. If A and A' share relevant features, and A' is known to cause B, then by analogy, A might be assumed to cause B as well.

These forms of evidence are inherently speculative and typically serve as initial heuristics to guide hypothesis generation or further empirical testing.

3.1.2. Contrastive Evidence

Contrastive evidence draws on observed differences in outcomes across varying conditions. Such arguments typically follow the structure:

> "A caused B because B occurs under A but not under $\neg A$."

We identify several sub-types of contrastive evidence:

Statistical covariation: A measurable difference in outcomes is observed between groups or conditions, and this difference persists even after adjusting for potential confounding variables. The contrast is interpreted as supporting a causal role for the varying factor.

Before-after comparison: An intervention or change is introduced, and a corresponding shift in outcomes is observed. If other factors remain stable, the contrast in outcome is attributed to the intervention.

Controlled experiment: All conditions are held constant except for the variable of interest. A consistent difference in outcome across conditions is then attributed to the manipulated variable.

3.1.3. Causal Explanation

Causal explanations articulate a mechanism that connects the cause to the effect. These arguments are typically stronger than purely correlational or contrastive forms due to their explanatory depth. They are often structured as:

> "A causes B because A initiates a sequence of intermediate steps leading to B."

We distinguish several sub-types of causal explanation:

Mechanistic explanation: The argument identifies a specific sequence of processes or interactions through which the cause produces the effect. This often involves reference to known physical, computational, or biological mechanisms.

Elimination of alternatives: A causal claim is supported by ruling out other plausible explanations. If the observed effect coincides only with changes in A, and other variables are held constant, A is inferred to be the cause.

Typicality of effect: The observed outcome matches the expected pattern associated with similar causes in comparable contexts. This reinforces the plausibility of the proposed mechanism.

3.2. Arguments Against Causal Claims

Arguments challenging causal claims fall into two categories: those questioning the plausibility of the causal relation and those attacking the logical structure or sufficiency of the justification.

3.2.1. Plausibility Challenges

These arguments suggest that the proposed causal link is implausible in light of available evidence. We distinguish several sub-types:

Wrong temporal order: The effect is observed before the supposed cause. Since causes must precede their effects, this undermines the causal interpretation.

No connection: The proposed cause and effect belong to unrelated domains, or no plausible pathway links them. Without a credible mechanism, the causal claim lacks support.

Free decision: The outcome results from an independent choice or intervention that is not causally determined by the proposed factor. The cause is incidental rather than explanatory.

Insufficient cause: The proposed factor occurs without reliably producing the effect. This suggests that it alone cannot account for the outcome and may require additional conditions.

Unnecessary cause: The effect can be fully explained by other causes. The proposed factor is therefore not needed to account for the outcome, weakening its causal relevance.

3.2.2. Logical Objections

Logical objections target the inferential structure of a causal argument, highlighting weak reasoning or offering superior alternatives. We distinguish several sub-types:

Alternative cause: The observed association between the proposed cause and effect can be better explained by a third, unaccounted-for variable that influences both.

Post hoc fallacy: The argument infers causality merely from temporal succession—assuming that because B followed A, A must have caused B—without further justification.

Low statistical support: The effect is observed inconsistently or weakly across instances. A low base rate or limited correlation challenges the robustness of the causal claim.

Anecdotal evidence: The argument relies on a single or highly atypical case, which is insufficient for generalisation and vulnerable to noise or confounding.

Unknown mechanism: No explanation is given for how the proposed cause leads to the effect. Without a plausible mechanism, the claim remains speculative.

3.2.3. Qualifying Causal Claims

Some arguments do not reject the existence of a causal link but instead refine or constrain its interpretation. These qualifications help clarify the nature, strength, or context of the causal relationship. We distinguish several sub-types:

Partial cause: The proposed factor contributes to the outcome but is not the sole or primary cause. Other influences play a more significant role.

Indirect cause: The effect arises through a chain of intermediate steps rather than a direct influence. The causal link is mediated by other variables.

Common cause: Both the proposed cause and the observed effect result from a shared underlying factor. The deeper cause lies elsewhere.

Interaction: The proposed factor produces the effect only in combination with other conditions. On its own, it may have little or no causal impact.

Reversed causality: The direction of influence is the opposite of what is claimed. What is presented as the cause is actually a response to the effect.

Accidental cause: The observed causal link is coincidental, resulting from an unrelated or unanticipated event that occurred simultaneously.

Listing 1 Prompt Template: Argue in Favour of a Causal Link

```
Given the causal claim: [INSERT CAUSAL CLAIM],
produce a structured argument *in support* of this claim. Choose only one among the
    following argumentation strategies:

1. Circumstantial Evidence:
  - Provide at least one argument using spatio-temporal contiguity, repeated co-
    occurrence, or similarity to known causes.

2. Contrastive Evidence:
  - Provide at least one argument based on statistical covariation, before-after
    comparison, or a controlled experiment.

3. Causal Explanation:
  - Provide at least one mechanistic explanation, argument from no alternative
    explanation, or an argument based on typical effect.

Each argument must:
- Clearly state which macro-family and specific subtype it belongs to.
- Be logically self-contained and persuasive.
- Refer explicitly to the observed or hypothesised phenomena.
- Be concise.
```

Listing 2 Prompt Template: Argue Against a Causal Link

```
Given the causal claim: [INSERT CAUSAL CLAIM],
produce a structured argument *against* this claim. Choose only one among the
    following argumentation strategies:

1. Plausibility Challenges:
  - Include at least one argument based on wrong temporal order, no plausible
    connection, free decision,
    insufficient cause, or unnecessary cause.

2. Logical Objections:
  - Include at least one argument from alternative cause, post hoc fallacy, low
    statistical support,
    anecdotal evidence, or unknown mechanism.

3. Qualifying Claims:
  - Optionally include partial, indirect, common cause, interaction, reversed
    causality, or accidental cause qualifications.

Each argument must:
- Identify the macro-family and the specific argumentation subtype.
- Be articulated as a counterpoint to a potential or actual supporting argument.
- Indicate whether the causal claim is to be rejected, weakened, or reformulated.
- Be concise.
```

4. Argumentative Evaluation of Tentative Causal Links via Language Models

Once tentative causal relationships have been identified using structure learning algorithms such as the PC algorithm (see Section 2), we proceed to critically evaluate these links using argumentative reasoning. To this end, we leverage large language models (LLMs) to generate structured arguments

Listing 3 Prompt Template: Final Judgement on a Causal Link

```
You are given two arguments regarding the same causal claim.

Causal Claim:
[INSERT CAUSAL CLAIM]

Argument in Favour:
[INSERT PRO-CASUAL ARGUMENT HERE]

Argument Against:
[INSERT COUNTER-ARGUMENT HERE]

Task:
Evaluate the overall credibility of the causal claim based on the two arguments.
    Your judgment must:

1. Reference the strength and relevance of the macro-family and subtype for each
    argument.
2. Indicate whether the causal claim should be:
  - Accepted as likely
  - Tentatively accepted with caveats
  - Undecided or requiring more evidence
  - Rejected
3. Justify your decision with explicit reference to the comparative argumentative
    strength (\eg directness, generalisability, mechanistic plausibility,
    alternative explanations).
4. Avoid introducing new arguments - focus only on evaluating the two provided.

Output format:
- Summary judgment (one sentence)
- Justification (3-5 sentences)
```

in favour of or *against* each causal claim, based on the typology of causal argumentation patterns introduced earlier (cf. Section 3 and [10, 11]).

The prompts we employ are explicitly designed to elicit responses that map onto macro-families of causal argumentation — *circumstantial evidence*, *contrastive evidence*, and *causal explanation* for supporting claims, and *plausibility challenges*, *logical objections*, and *qualifications* for opposing ones. These prompt templates are provided in Listings 1 and 2.

For instance, suppose the PC algorithm outputs the link `batch_size` → `power`. We then instantiate the favour prompt (see Listing 1) as:

> "Given the causal claim: 'Large batch sizes cause increased GPU power consumption', produce a structured argument *in support* of this claim..."

This would result in the LLM producing arguments such as:

- **Circumstantial evidence (repeated co-occurrence):** "In over a dozen training configurations, large batch sizes were consistently associated with high energy usage."
- **Contrastive evidence (controlled experiment):** "When only the batch size was varied, all other factors held constant, energy draw increased with larger batches."
- **Causal explanation (mechanistic):** "Larger batch sizes lead to greater parallelisation, saturating GPU resources and increasing power draw."

This argumentative evaluation serves two purposes. First, it adds interpretability and justifiability to data-driven causal claims. Second, it exposes weak or unsupported inferences by juxtaposing strong and weak forms of justification within the same causal hypothesis.

To complement the generation of individual arguments for and against causal hypotheses, we introduce a final reasoning stage where an LLM is prompted to adjudicate between competing views. This dialectical evaluation phase simulates the structure of critical discussion in causal reasoning. Given one argument supporting a causal link and one refuting it, the model is tasked with determining the overall plausibility of the claim. The prompt is designed to ensure that the response explicitly references both the argumentative structure and the relative strength of the opposing macro-families. The complete template is shown in Listing 3.

This stage enforces a structured comparison. The LLM is instructed to identify the macro-family and subtype of each input argument, evaluate their logical and evidential strength, and conclude with one of four options: the causal claim is (i) accepted as likely, (ii) tentatively accepted with caveats, (iii) undecided and requiring more evidence, or (iv) rejected. The rationale behind this design is to integrate argumentative quality into the causal assessment pipeline, preventing premature commitment to claims generated by data-driven discovery alone.

Consider again the hypothesis that "large batch sizes cause increased GPU power consumption". Suppose we input the following arguments:

- **Argument in favour (Contrastive evidence - controlled experiment):** "When all training parameters were kept constant except for batch size, power draw increased with larger batches."
- **Argument against (Logical objection - alternative cause):** "Both batch size and energy usage correlate with model complexity, which is the more likely common cause."

The LLM, following the prompt in Listing 3, may respond:

> *Summary judgment: Tentatively accepted with caveats.*
>
> *Justification: The argument in favour is based on a controlled experiment, which provides strong contrastive evidence and isolates batch size as a plausible causal factor. The counter-argument rightly identifies a potential confound (model complexity), but does not negate the observed contrastive pattern. While further control for model size is advisable, the causal link is tentatively supported.*

5. Preliminary Results

5.1. Dataset: The BUTTER-E Benchmark

To support the empirical evaluation of causal discovery methods in energy-sensitive machine learning pipelines, we employ the **BUTTER-E dataset** [17, 18], a large-scale benchmark of real-world energy consumption measurements associated with Deep Neural Network (DNN) training. Specifically, BUTTER-E extends the *primary sweep* of the original BUTTER dataset by adding detailed energy usage data collected on both CPU and GPU hardware. The resulting dataset comprises over 41,000 distinct training runs of fully connected multi-layer perceptrons (MLPs), each corresponding to a unique configuration of hyperparameters.

The training runs in BUTTER-E span a diverse space of model sizes, shapes, and depths. The number of trainable parameters ranges from 2^5 to 2^{25}, distributed across 2 to 20 hidden layers. Eight architectural shapes were explored, including:

- `rectangle` and `rectangle_residual` (uniform width, with or without residual connections),
- `trapezoid` (width decreases linearly with depth),
- `exponential` (exponential decay in width),

- and `wide_first_`n`x` configurations with $n \in \{2, 4, 8, 16\}$, where the first hidden layer is n times wider than the subsequent layers.

The training datasets were selected from the Penn Machine Learning Benchmark (PMLB) repository [19, 20], and include: `201_pol`, `294_satellite_image`, `529_pollen`, `537_houses`, `adult`, `banana`, `connect_4`, `MNIST`, `nursery`, `sleep`, `splice`, and `wine_quality_white`.

The energy measurements were obtained by re-executing the primary sweep of BUTTER on the **Eagle** high-performance computing (HPC) cluster at NREL. Each run was assigned to a dedicated CPU or dual-GPU compute node. CPU training was executed using Intel's OneDNN-accelerated TensorFlow, while GPU training used cuDNN and TensorFlow with NVIDIA-recommended settings. Instantaneous power consumption was recorded at one-minute intervals using the Hewlett-Packard Enterprise Integrated Lights-Out (iLO) monitoring system embedded in each node.

Rather than treating all available features indiscriminately, we adopted a selective filtering process based on substantive criteria. Features were retained if they were a) plausibly causal in their relationship to energy usage (*e.g.*, depth, width, optimizer), b) empirically meaningful across the run population (*e.g.*, dataset, GPU usage), and c) measurable prior to training (ensuring suitability for forward causal inference). This led to the retention of the following fields: `size_x`, `depth_x`, `shape_x`, `dataset_x`, `learning_rate_x`, `batch_size_x`, `optimizer_x`, `is_gpu_x`, and the target variable `power`.

Crucially, this design preserves the interpretability and modularity of the resulting dataset. The extracted subset corresponds to a well-defined scientific object: a tabular representation of experimental factors and outcomes, abstracted away from runtime-specific details or monitoring artefacts. Each row is interpretable as a complete experimental unit, amenable to statistical modelling, formal argumentation, or simulation-based what-if analysis. By retaining only the subset of features that are non-redundant and structurally important, we minimise the risk of collider bias and improve the tractability of causal graph learning procedures.

5.2. Causal Graph Inferred via the PC Algorithm

To investigate potential causal dependencies between architectural, training, and execution-related variables, we employed the PC algorithm (Section 2). The method was applied to the harmonised subset of the BUTTER-E dataset containing architectural parameters (*e.g.*, depth, size, shape), training configurations (*e.g.*, batch size, learning rate, optimizer), execution hardware (*e.g.*, GPU usage), and outcome measures (*e.g.*, power consumption). The PC algorithm was configured with a significance threshold $\alpha = 0.001$, a maximum conditioning set size of $c = 2$, and Fisher's Z-test for conditional independence. Column names were sanitised by removing suffixes such as `_x` for clarity.

The PC algorithm operates in two main phases: the skeleton identification phase and the edge orientation phase. In the first phase, a fully connected undirected graph is pruned by iteratively testing for conditional independence between pairs of variables, conditioning on subsets of increasing size. In the second phase, the remaining edges are oriented using the rules of causal sufficiency, v-structure identification, and the application of Meek's rules [21]. The significance level $\alpha = 0.001$ defines the tolerance for Type I error in independence tests, enforcing a conservative edge removal strategy.

The independence test used in our pipeline is Fisher's Z-test, which evaluates whether two variables X and Y are conditionally independent given a set Z by testing whether their partial correlation $\rho_{XY|Z}$ is significantly different from zero.

The inferred graph structure, shown in Figure 1, displays several dense regions of influence, notably around the `optimizer`, `depth`, and `batch_size` nodes. The `optimizer` node emerges as a global influencer, with causal edges toward almost all other configuration variables, including `size`, `depth`, `shape`, `learning_rate`, and even hardware utilisation (`is_gpu`). This pattern may reflect the fact that optimizers are typically selected early in the training pipeline, and that this choice often constrains or influences subsequent design decisions.

The node `depth` is positioned as an intermediate confounder, affecting both `size` and `batch_size`. This is consistent with the fact that deeper networks tend to require larger parameter counts and more

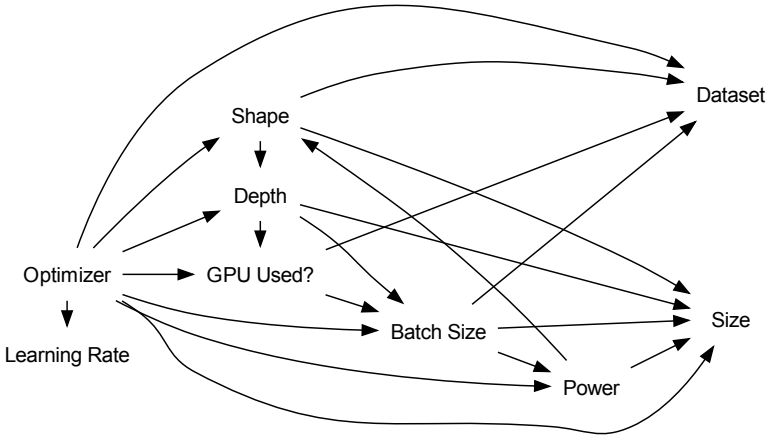

Figure 1: CPDAG resulting from the PC algorithm.

aggressive batching strategies for convergence. Similarly, `shape` is shown to influence not just network size and depth but also the dataset, suggesting that architectural templates are adapted to the complexity or dimensionality of the task.

The inferred structure reveals not only direct influence pathways but also points of intervention. For instance, altering the `optimizer` may simultaneously shift multiple architectural choices and downstream power consumption. Similarly, `batch_size` appears as a critical mediator of both model complexity and energy cost.

Interestingly, `power` is not only a sink (with arrows from `optimizer` and `batch_size`) but also acts as a source node for `size` and `shape`. While this may initially appear paradoxical, since energy consumption is usually treated as an outcome, it may reflect feedback loops during neural architecture search or early stopping criteria that prune high-power configurations, thereby shaping subsequent architecture choices. Alternatively, this pattern could signal violations of faithfulness or the presence of latent confounders such as runtime hardware throttling or dynamic learning rate adjustment. We will focus on this link further in Section 5.3

5.3. Critical Evaluation of LLM Output on Causal Claim

To demonstrate the methodology and the type of results that can be obtained through argumentative evaluation of causal links, we focus in this paper on a single, non-obvious causal claim: "Power consumption during training of the weights of a neural network causes the engineering choice of the size of the neural network." This example was selected intentionally for its subtle and context-dependent nature, making it a suitable candidate to test the depth and coherence of reasoning produced by a large language model. The evaluation was conducted using GPT-4o (`gpt-4o`, release date May 13, 2024), which was prompted to generate three outputs: an argument in favour of the causal link (Listing 4), an argument against it (Listing 5), and a final judgment synthesising both perspectives (Listing 6).

The pro-causal argument in Listing 4 aligns with the *Causal Explanation* macro-family, and more specifically with the *Mechanistic Explanation* subtype. It presents a coherent engineering feedback loop in which anticipated or previously observed energy demands during training act as constraints that shape the design of future architectures. This reflects practices common in real-world machine learning development—particularly under conditions of limited compute or sustainability requirements. The argument implicitly assumes that such considerations are consistently integrated into design workflows, which may not universally hold across experimental or academic settings.

The counterargument in Listing 5 adopts a *Plausibility Challenge* stance, relying on the *Wrong*

Listing 4 LLM-generated argument in favour of the causal claim: power consumption during training causes engineering decisions on network size.

```
Macro-family: Causal Explanation
Subtype: Mechanistic Explanation

Structured Argument in Support:

The claim that power consumption during training of the weight of a neural network
    causes the engineering choice of the size of the neural network can be
    supported by a mechanistic explanation grounded in how neural network
    development proceeds in practice.

During training, a neural network's energy consumption increases with both the
    number of parameters and the computational workload (\eg forward and backward
    passes, gradient updates). Larger models demand more GPU/TPU resources, leading
     to longer training times and higher energy costs. This power consumption is
    directly measurable and often becomes a bottleneck when scaling models,
    especially in settings with fixed energy budgets, sustainability targets, or
    hardware limitations.

Engineers, aware of these resource constraints, often incorporate them into the
    design phase as constraints on model architecture. For example, training a 10-
    billion parameter model on a modest compute cluster may be prohibitively
    expensive in terms of power. Therefore, engineers reduce the model size
    preemptively to meet acceptable power consumption levels during training.

This mechanism, where observable power demands during training feed back into
    design constraints, explains how power consumption causally determines the
    engineering decision of network size. The process aligns with standard
    engineering feedback loops where operational resource costs shape future design
     choices.
```

Listing 5 LLM-generated counterargument: the causal claim fails due to incorrect temporal ordering.

```
Macro-family: Plausibility Challenges
Subtype: Wrong Temporal Order

Counterpoint:
A supporting argument might suggest that higher power consumption during training
    compels engineers to reduce network size. However, in practice, the size of the
     neural network is determined during the model design phase, which precedes any
     training or associated power consumption. Therefore, the temporal order does
    not support causality.

Effect on Causal Claim: Rejected
```

Temporal Order subtype. It highlights that model size decisions necessarily precede training, and therefore post-training power consumption cannot causally influence architectural choices. While formally valid, this argument overlooks the iterative nature of engineering processes, where anticipated power usage—based on prior experience or simulations—can influence decisions made before training actually begins. As such, it provides a useful but incomplete critique.

Finally, the judgment in Listing 6 characterises the causal claim as "tentatively accepted with caveats," offering a balanced assessment. It recognises the explanatory strength and practical relevance of the mechanistic argument, while acknowledging the validity of the temporal critique. The model

Listing 6 LLM-generated final judgment on the causal claim, balancing both sides.

```
Summary judgment:
Tentatively accepted with caveats.

Justification:
The supporting argument employs a Causal Explanation within the Mechanistic
    Explanation subtype, which is well-suited for establishing a plausible internal
     process linking observed power consumption during training to design decisions.
     It offers a concrete mechanism by which engineers anticipate resource
    constraints based on historical or expected consumption profiles, thereby
    integrating them into the model design process. While the counterargument
    identifies a Plausibility Challenge based on Wrong Temporal Order, it treats "
    training power consumption" as a post hoc observation rather than an
    anticipated constraint, which may not fully reflect iterative engineering
    workflows where anticipated training costs inform design. The mechanistic
    account carries more explanatory weight due to its grounding in practical
    engineering constraints and feedback mechanisms, although it does not fully
    address whether the causal link is universally applicable. Therefore, the claim
     is tentatively accepted, pending further clarification of how often and
    explicitly such power considerations directly determine network size.
```

appropriately frames the causality as context-dependent and suggests that further empirical investigation would be necessary to determine how routinely power constraints explicitly guide network sizing decisions.

6. Limitations

While the results presented in this paper provide promising evidence for the feasibility of using large language models (LLMs) to generate and evaluate causal arguments, several limitations remain that point to directions for future research.

First, the methodology relies on a single prompt format for each task — namely, for generating arguments in favour of and against a causal claim, and for producing a final judgment. Although these prompts were carefully crafted to align with established principles from argumentation theory, this design choice inevitably constrains the expressive richness and adaptability of the generated reasoning. A more comprehensive study of prompt design is needed. This should include the exploration of alternative phrasings and structured prompting schemes that instantiate different argumentation strategies, such as Walton's argumentation schemes [22], abductive reasoning patterns, or counterfactual-based templates. These variations would allow the system to capture a broader typology of causal reasoning styles.

In addition, adopting few-shot learning strategies using curated examples of high-quality causal arguments could enhance both the consistency and epistemic soundness of the generated content. Beyond this, one may envision prompts that produce structured outputs in agentic form — where each argument is tagged with a scheme type, source of support, and confidence level — such that they can be automatically passed to formal solvers grounded in computational argumentation theory. These could include Dung-style abstract argumentation frameworks [23], structured argumentation formalisms like ABA [24] or ASPIC+ [25], or probabilistic extensions that accommodate uncertainty in weights or justifiability [26]. In such settings, the LLM's role would be to simulate a deliberative agent [27] capable of producing argumentative material in a machine-readable form, supporting downstream computational evaluation leveraging existing argumentation solvers [28].

Second, the current pipeline assumes that all necessary domain knowledge is already embedded within the language model. In our case, GPT-4o was used without access to any external knowledge retrieval mechanism. While this model has demonstrated strong capabilities in handling technical content, including machine learning literature — see, *e.g.*, [29] for general capabilities and [30] for LLMs'

performance on domain-specific tasks — the absence of a domain-aware knowledge management system limits robustness and interpretability. In practice, real-world causal evaluation may benefit from the integration of Retrieval-Augmented Generation (RAG) pipelines [31, 32, 12] that allow the LLM to reference verified domain knowledge, documentation, or empirical findings. This would also enable traceable citation practices and mitigate hallucinations.

Third, the same language model was used for both the generation of arguments and their evaluation. While this simplifies the experimental setup, it raises concerns about internal consistency biases and limited critical distance. A more robust approach may involve using distinct models for generation and judgment — potentially including domain-specific models for technical content generation and models fine-tuned specifically for argumentative coherence, logical soundness, or critical reasoning. Such model pluralism would allow for cross-verification and better reflect a multi-agent argumentative process.

7. Conclusion and Future Work

This study proposed an initial framework for enhancing data-driven causal discovery through argumentative evaluation, leveraging large language models (LLMs) to generate and assess natural language justifications for inferred causal links. By framing each link as open to dialectical scrutiny, the approach aims to make causal inference more transparent and interpretable. While the results are promising, several limitations highlight avenues for future work. These include exploring diverse prompt designs — such as few-shot prompting with exemplary arguments — to improve output quality; integrating Retrieval-Augmented Generation (RAG) to supplement model knowledge with curated external sources; and adopting a modular architecture that separates generation and evaluation roles across specialised models to mitigate coherence bias. These directions support the development of a hybrid causal inference pipeline that is both statistically rigorous and discursively robust.

Acknowledgments

This work was supported by the EU NEXTGENERATIONEU program within the PNRR Future Artificial Intelligence – FAIR project (PE0000013, CUP H23C22000860006), Objective 10: Abstract Argumentation for Knowledge Representation and Reasoning, specifically by the project Argumentation for Informed Decisions with Applications to Energy Consumption in Computing – AIDECC (CUP D53C24000530001). This work was supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU, specifically by the project NEACD: Neurosymbolic Enhanced Active Cyber Defence (CUP J33C22002810001). This work was supported by project ACRE (AI-Based Causality and Reasoning for Deceptive Assets - 2022EP2L7H) and xInternet (eXplainable Internet - 20225CETN9) projects - funded by European Union -Next Generation EU within the PRIN 2022 program (D.D. 104 - 02/02/2022 Ministero dell'Università e della Ricerca). The work was partially supported by the European Office of Aerospace Research & Development and the Air Force Office of Scientific Research under award number FA8655-22-1-7017 and by the US DEVCOM Army Research Laboratory (ARL) under Cooperative Agreement #W911NF2220243. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States government.

Declaration on Generative AI

Some of the results presented in this paper were obtained using ChatGPT-4o as part of the research process. The authors also used ChatGPT-4o for grammar and spelling checks, as well as for generating and refining portions of the text. All outputs from the tool were reviewed and edited by the authors as necessary, and the authors take full responsibility for the accuracy, integrity, and final presentation of the publication's content.

References

[1] J. Chai, H. Zeng, A. Li, E. W. Ngai, Deep learning in computer vision: A critical review of emerging techniques and application scenarios, Machine Learning with Applications 6 (2021) 100134.

[2] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, S. Azam, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access (2024).

[3] R. Zheng, L. Qu, B. Cui, Y. Shi, H. Yin, AutoML for deep recommender systems: A survey, ACM Transactions on Information Systems 41 (2023) 1–38.

[4] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, S. Watanabe, End-to-end speech recognition: A survey, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023).

[5] D. Geißler, B. Zhou, M. Liu, S. Suh, P. Lukowicz, The power of training: How different neural network setups influence the energy demand, in: International Conference on Architecture of Computing Systems, Springer, 2024, pp. 33–47.

[6] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green ai, Communications of the ACM 63 (2020) 54–63.

[7] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, arXiv arXiv:1906.02243 (2019).

[8] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Kording, C. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, Y. Bengio, Tackling climate change with machine learning, arXiv arXiv:1906.05433 (2019).

[9] N. Jones, et al., How to stop data centres from gobbling up the world's electricity, Nature 561 (2018) 163–166.

[10] U. Oestermeier, F. W. Hesse, Verbal and visual causal arguments, Cognition 75 (2000) 65–104.

[11] A. Bochman, F. Cerutti, T. Rienstra, Causation and argumentation, Journal of Applied Logics 12 (2025) 713–786.

[12] D. Jurafsky, J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd ed., 2025. URL: https://web.stanford.edu/~jurafsky/slp3/, online manuscript released January 12, 2025.

[13] J. Pearl, Causal diagrams for empirical research, Biometrika 82 (1995) 669–688. Publisher: Oxford University Press.

[14] J. Pearl, Causality: models, reasoning, and inference, Cambridge University Press, 2009.

[15] J. Pearl, M. Glymour, N. P. Jewell, Causal inference in statistics: A primer, John Wiley & Sons, 2016.

[16] P. Spirtes, C. Glymour, R. Scheines, Causation, prediction, and search, MIT press, 2001.

[17] C. Tripp, J. Perr-Sauer, E. Bensen, J. Gafur, A. Nag, A. Purkayastha, Butter-e - energy consumption data for the butter empirical deep learning dataset, Open Energy Data Initiative (OEDI), National Renewable Energy Laboratory, https://doi.org/10.25984/2329316, 2022.

[18] C. E. Tripp, J. Perr-Sauer, J. Gafur, A. Nag, A. Purkayastha, S. Zisman, E. A. Bensen, Measuring the energy consumption and efficiency of deep neural networks: An empirical analysis and design recommendations, arXiv preprint arXiv:2403.08151 (2024).

[19] R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz, J. H. Moore, Pmlb: a large benchmark suite for machine learning evaluation and comparison, BioData Mining 10 (2017) 36.

[20] J. D. Romano, T. T. Le, W. La Cava, J. T. Gregg, D. J. Goldberg, P. Chakraborty, N. L. Ray, D. Himmelstein, W. Fu, J. H. Moore, Pmlb v1.0: an open source dataset collection for benchmarking machine learning methods, arXiv preprint arXiv:2012.00058v2 (2021).

[21] C. Meek, Causal inference and causal explanation with background knowledge, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995), 1995.

[22] D. N. Walton, C. Reed, F. Macagno, Argumentation Schemes, Cambridge University Press, NY, 2008.

[23] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial intelligence 77 (1995) 321–357.

[24] A. Bondarenko, F. Toni, R. A. Kowalski, An assumption-based framework for non-monotonic reasoning., in: LPNMR, volume 93, 1993, pp. 171–189.
[25] S. Modgil, H. Prakken, A general account of argumentation with preferences, Artificial Intelligence 195 (2013) 361–397.
[26] A. Hunter, A probabilistic approach to modelling uncertain logical arguments, International Journal of Approximate Reasoning 54 (2013) 47–81.
[27] D. B. Acharya, K. Kuppan, B. Divya, Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey, IEEE Access (2025).
[28] F. Cerutti, S. A. Gaggl, M. Thimm, J. Wallner, Foundations of implementations for formal argumentation, Journal of Applied Logics 4 (2017) 2623–2705.
[29] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).
[30] P. Törnberg, Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning, arXiv preprint arXiv:2304.06588 (2023).
[31] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.
[32] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A survey on rag meeting llms: Towards retrieval-augmented large language models, in: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 6491–6501.