Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.076507 Published Online: 27 February 2026

ARTICLE

Semantic Causality Evaluation of Correlation Analysis Utilizing Large Language Models

Adam Dudáš*

Department of Computer Science, Faculty of Natural Sciences, Matej Bel University, Tajovského 40, Banská Bystrica, Slovakia

*Corresponding Author: Adam Dudáš. Email: adam.dudas@umb.sk

Received: 21 November 2025; Accepted: 02 February 2026

ABSTRACT: It is known that correlation does not imply causality. Some relationships identified in the analysis of data are coincidental or unknown, and some are produced by real-world causality of the situation, which is problematic, since there is a need to differentiate between these two scenarios. Until recently, the proper–semantic–causality of the relationship could have been determined only by human experts from the area of expertise of the studied data. This has changed with the advance of large language models, which are often utilized as surrogates for such human experts, making the process automated and readily available to all data analysts. This motivates the main objective of this work, which is to introduce the design and implementation of a large language model-based semantic causality evaluator based on correlation analysis, together with its visual analysis model called Causal heatmap. After the implementation itself, the model is evaluated from the point of view of the quality of the visual model, from the point of view of the quality of causal evaluation based on large language models, and from the point of view of comparative analysis, while the results reached in the study highlight the usability of large language models in the task and the potential of the proposed approach in the analysis of unknown datasets. The results of the experimental evaluation demonstrate the usefulness of the Causal heatmap method, supported by the evident highlighting of interesting relationships, while suppressing irrelevant ones.

KEYWORDS: Correlation; causality; correlation analysis; large language models; visualization

1 Introduction

Correlation analysis is frequently used in a broader scope of analysis of data to identify interesting functional relationships in the studied data. However, these relationships can be reflective of reality, where the two studied attributes influence each other's values—making such a relationship causal—but also can only be present in the specific measured data instances. Correlation and causality are closely related concepts, with correlation analysis often serving as a starting point for investigating whether deeper causal mechanisms are present [1].

Recently, large language models have been used as a system component with the potential to act as human expert surrogates, capable of bringing broad, real-world knowledge into analytical workflows [2,3]. When considering the semantic evaluation of causality of relationships identified in data, causality assessment requires such human experts from the domain of the studied data to judge whether an observed relationship is plausibly causal or coincidental/unknown. This dependence on expert judgment can be time-consuming, significantly limits the scalability of the task, and depends on experts for areas of expertise, where there might be only a few of them.

This motivated the core idea of the presented study—the application of large language models in the task of semantic evaluation of the real-world causality of relationships identified via standard correlation analysis. Unlike conventional causal methods that rely purely on statistical models and tests, this approach leverages the online nature of large language models, making them able to assess the real-world familiarity and plausibility of relationships even in previously unseen datasets, offering a complementary approach to traditional expert- or statistics-driven causal analysis. In this way, the study contributes a novel visual analysis method that uses large language models to semantically evaluate the causal plausibility of relationships initially discovered through correlation analysis.

Hence, objectives of the work can be summarized into the following points:

- Design of procedure in which large language models are used to semantically evaluate the causality of relationships identified in correlation analysis and produce novel, weighted coefficient value, which reflects this evaluation.
- Design of a visual analysis method called Causal heatmap to present the results of the causal evaluation process in a visual way.
- Implementation of the causal evaluator and Causal heatmap in *Python* language to produce an open-source solution for semantic causal analysis of relationships in data.
- Experimental evaluation of the proposed model conducted on the combination of two datasets, two different causal evaluation problems, and three different evaluatory points of view—the quality of the visual model, the quality of semantic causal evaluation based on large language models, and comparative analysis of the method with conventional approaches.

Other than the introduction itself, the presented work consists of four main sections of text. Section 2 briefly examines scientific literature related to the objectives of the work, focusing on the use of large language models as surrogates for human experts. In Section 3, the design for the proposed causal evaluation workflow and visualization model are presented. The implementation and experimental evaluation of the model are detailed in Section 4, while the work concludes with Section 5, addressing advantages, disadvantages, and possible future work ideas related to the studied area.

2 Related Works

As stated in the introductory section of this text, this work aims towards the use of large language models as surrogates for human experts, which would identify known causality and/or novel relationships in the studied data. The utilization of large language models as a human expert surrogate is well documented in several research outputs, out of which this section of the work presents only a few modern results.

In [4], authors focus on the application of large language models to process the textual representation of molecules studied in various areas of molecular science. Since the previous language model-based approaches to the task proved ineffective, the authors introduce GIT-Mol—a multi-modal large language model capable of working with visual and textual information—together with a novel architecture called GIT-Former. These novel contributions are reported to increase the accuracy of molecular property prediction by up to 10% and the quality of molecule generation by up to 20%.

The authors of [5] present the use of a large language model in the context of energetics—specifically, the model is used to determine the partial tripping of distributed energy resources based on the properties of interest. In the study, the authors utilize a BERT-based approach to streamline the fault information into tokenized input, which, on one hand, reduces the complexity of machine learning models needed for the selected task and, on the other, demonstrates the high quality of performance on limited sets of data.

The research presented in [6] focuses on the human interaction with radio map generation and wireless network planning, which often requires complex manual operations, limiting the use of automation in the problem. To counter this constraint, the authors of the work propose a large language model-based solution, in which the model autonomously generates radio maps and facilitates wireless network planning for specified areas, hence minimizing the need for extensive human-computer interaction. The reached results prove that the large language model utilization in the task reduces the amount of manual operations needed, while achieving enhanced coverage and signal-to-interference-noise ratio.

On the other hand, the research presented in [7] aims to apply large language models for remote sensing data. The authors of the study introduce three novel additions to the area—a large-scale remote sensing instruction tuning dataset, a SkyEyeGPT language model specifically designed for remote sensing, and a two-stage tuning method for the enhancement of instruction-following of the large language model in the selected task. Experimental results show that the combination of the three novel outputs produces superior outputs to the most conventional modern approaches.

The work presented in this study encompasses diagnostic analysis of data based on the correlation analysis and the determination of causality of relationships identified in it based on a large language model-based evaluator. This objective is a form of continuation of the previous research presented in [8,9] in which a correlation matrix was linguistically described by a large language model as shown in Fig. 1. In this previous approach, a correlation matrix is constructed, then a large language model generates a brief linguistic description for each of the identified relationships, and lastly, the matrix is supplemented by diagnostic cards, which describe the relationships.

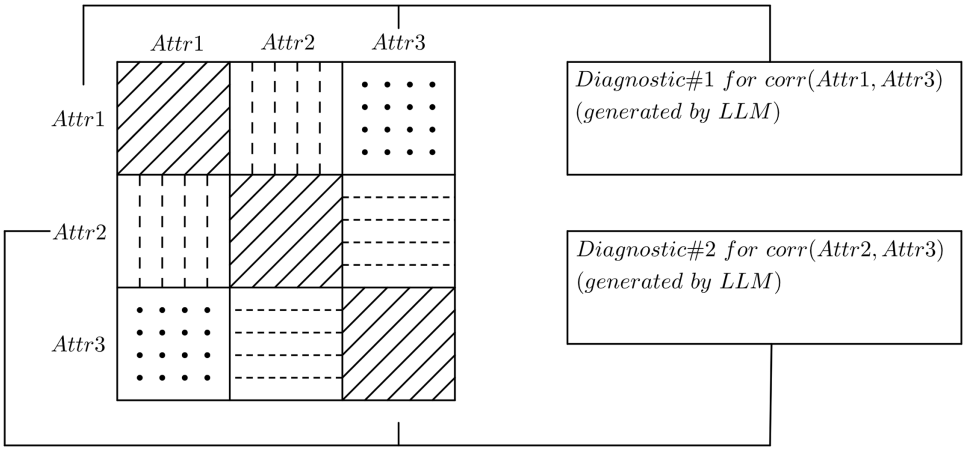

Figure 1: Schematic visualization of the approach presented in [8,9].

Since such a diagnostic analysis visualization of correlation matrices commonly contained mixed results and a high number of unnecessary descriptions, the results point to the need for the use of methods which highlight the causality or novel, unknown relationships in the correlation matrix. This motivates the utilization of large language models as causal evaluators of correlation analysis presented in the work detailed in this text.

3 Semantic Causal Analysis Utilizing Large Language Models

The original approach to semantic causal analysis of multidimensional datasets presented in this work is composed of the three basic building blocks—correlation analysis, large language models, and visualization.

In this section of the text, each of these elements of the method and their use in the causal analysis are described.

The foundation of the proposed model is correlation analysis, a statistical method for the identification of the strength and direction of functional relationships between pairs of attributes in a multidimensional dataset [10]. These relationships are described using correlation coefficients, which produce values from the $[-1, 1]$ interval—where the absolute value denotes the strength of the relationship between the measurements of studied attributes and the sign denotes the direction of this relationship. Naturally, the further the value of the correlation coefficient is from 0, the stronger the relationship itself.

In the conventional correlation analysis, two types of correlation coefficients are used—the Pearson correlation coefficient for the linear relationships, and the Spearman rank correlation coefficient for non-linear monotone relationships. The first of the mentioned correlation coefficients measured on attributes $attr_A$ and $attr_B$ is computed as follows [11]:

$$r(attr_A, attr_B) = \frac{\sum_{i=1}^{n}(attr_A i - \mu(attr_A))(attr_B i - \mu(attr_B))}{\sqrt{\sum_{i=1}^{n}(attr_A i - \mu(attr_A))^2}\sqrt{\sum_{i=1}^{n}(attr_B i - \mu(attr_B))^2}} \tag{1}$$

where $\mu(attr)$ is the mean value of the attribute $attr$ and n is the number of measurements of values for the considered attributes. However, since the Pearson correlation coefficient strongly depends on mean values of attributes, its values are significantly influenced by non-normal distributions and outliers.

These disadvantages motivate the use of the Spearman rank correlation coefficient computed as [12]:

$$\rho(attr_A, attr_B) = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{2}$$

where d denotes the difference between ranking of i-th measurements for the considered attributes. This ranking is assigned to individual values of an attribute in an ascending order; therefore, the specific values are abstracted from which makes the work with outliers and other than normal probabilistic distributions of attribute values possible.

Naturally, correlation between the values of two attributes can be the result of either real-world known causality of the relationship, or some other aspect—such as unknown or coincidental relationships. To identify the level of familiarity in linguistic, human-centric form, one can classify discovered relationships into five categories, such as:

$$relationship(attr_A, attr_B) = \begin{cases} known \\ less\ known \\ neutral \\ less\ unknown \\ unknown \end{cases} \tag{3}$$

Each of these categories can be assigned to the studied relationship between a pair of attributes by a human expert from the domain of the data or by their surrogate in an automized system, such as a large language model as proposed in this study. This evaluator of relationship familiarity can, then, use these linguistic values to produce a value for the familiarity level coefficient (ω) in two ways, based on the objective of the analysis:

- In the case, the objective of the data analysis is the identification of new relationships, the ω coefficient should increase for less known relationships and decrease for well-known relationships, such as:

$$\omega(attr_A, attr_B) = \begin{cases} 0.1, & \text{if the relationship is } known \\ 0.3, & \text{if the relationship is } less\ known \\ 0.5, & \text{if the relationship is } neutral \\ 0.7, & \text{if the relationship is } less\ unknown \\ 1, & \text{if the relationship is } unknown \end{cases} \tag{4}$$

- In such a case, when the objective of analysis is the causality identification, the ω values should be assigned in the opposite way:

$$\omega(attr_A, attr_B) = \begin{cases} 1, & \text{if the relationship is } known \\ 0.7, & \text{if the relationship is } less\ known \\ 0.5, & \text{if the relationship is } neutral \\ 0.3, & \text{if the relationship is } less\ unknown \\ 0.1, & \text{if the relationship is } unknown \end{cases} \tag{5}$$

After the determination of the familiarity level coefficient value for each of the attribute pairs, simple weighing of correlation values is computed as the semantic causal coefficient ($caus$):

$$caus(attr_A, attr_B) = \omega(attr_A, attr_B) \cdot corr(attr_A, attr_B) \tag{6}$$

When applied to the whole set of correlation coefficients of a dataset, the approach produces a causal matrix constructed as follows:

$$\begin{array}{cccccc} & \mathbf{attr_A} & \mathbf{attr_B} & \mathbf{attr_C} & \dots \\ \mathbf{attr_A} & 0 & \text{caus}(attr_A, attr_B) & \text{caus}(attr_A, attr_C) & \dots \\ \mathbf{attr_B} & \text{caus}(attr_B, attr_A) & 0 & \text{caus}(attr_B, attr_C) & \dots \\ \mathbf{attr_C} & \text{caus}(attr_C, attr_A) & \text{caus}(attr_C, attr_B) & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array} \tag{7}$$

This procedure should maintain the relationships of interest based on the objective of analysis, while suppressing the less interesting values (pushing them closer to 0), and not inflating values on the edge, which can be of interest, but also can be uninteresting. The whole presented procedure can be summarized in the schema in Fig. 2.

As can be seen in the figure, the last step of the proposed model consists of visualization of the causal matrix in the form of a causal heatmap, presented in Fig. 3. This visualization is composed of three basic sections—the main one being the heatmap itself (Fig. 3b), which contains the values of causal matrix visualized using color from the specified palette (Fig. 3a); the last part of the visualization (Fig. 3c) presents the sum of familiarity level coefficient ($\sum(\omega)$) visualized as bar graph for the purposes of analysis of familiarity levels related to individual attributes.

Therefore, the presented approach produces a visual model for the analysis of pairwise relationships in the studied data, which can be used as a part of diagnostic analysis of data and to identify novel knowledge from the data itself.

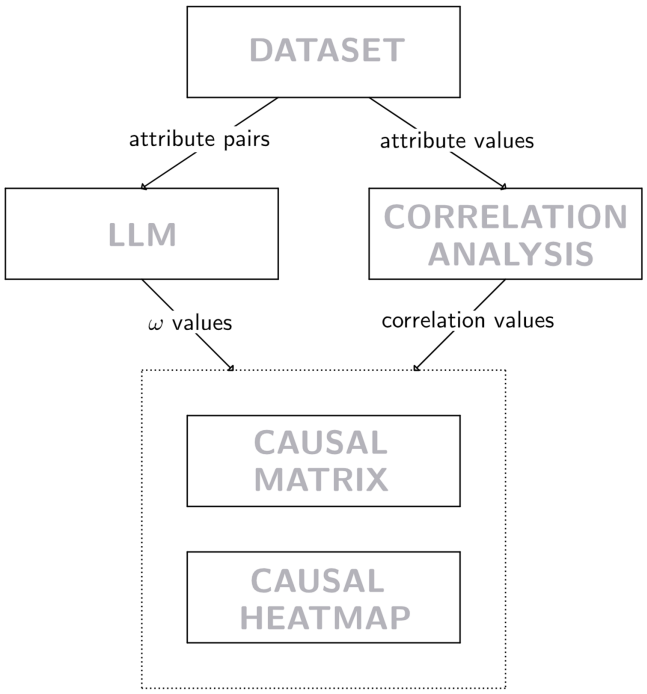

Figure 2: Schematic flowchart of the proposed approach.

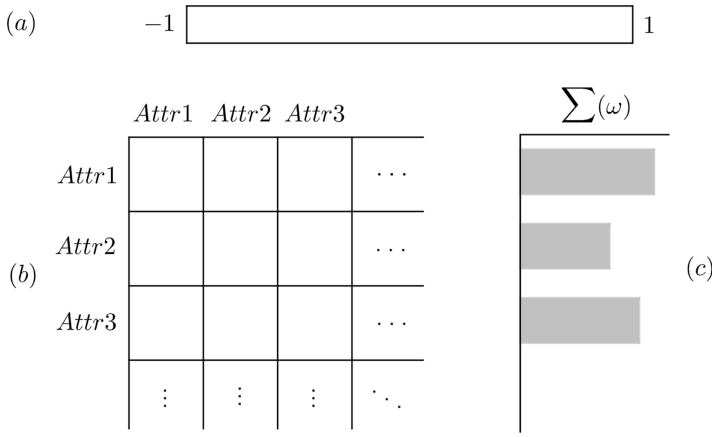

Figure 3: Schema of the Causal heatmap, where (***a***) denotes color scale used in the heatmap, (***b***) denotes the causal heatmap itself, and (***c***) visualizes the sum of ω values.

4 Evaluation of the Proposed Approach

The semantics-based causal evaluation model proposed in the previous section of this work was implemented in the *Python* language using several of the conventionally utilized packages for statistical analysis of data, such as *pandas* and *numpy*, and *matplotlib* and *seaborn* for visualization of data. To implement a large language model needed for the causal evaluation of relationships, the *Gemini* 2.5 *Flash* by Google was implemented through its API [13].

In regard to the language model prompting used in the implementation of the proposed model, several iterations were tested, while the following prompts led to the best performance of the system:

- Prompting for the problem of identification of novel (unknown) relationship:

```
Identify general familiarity of the relationship between pairs
of attributes:
{attributes}.
Evaluate these relationships as follows: known = 0.1 | moderately
known = 0.3 |
neutral = 0.5 | moderately unknown = 0.7 | unknown = 1.
The data come from the area of {area}.
Print only the pair title and the score.
```

- Prompting for the task of causality identification:

```
Identify general familiarity of the relationship between pairs
of attributes:
{attributes}.
Evaluate these relationships as follows: known = 1 | moderately
known = 0.7 |
neutral = 0.5 | moderately unknown = 0.3 | unknown = 0.1.
The data come from the area of {area}.
Print only the pair title and the score.
```

As seen in the final prompts implemented in the causal evaluator, some minor formatting of the language model output was conducted, together with one crucial element of the prompt—the addition of the domain (area) from which the data originates. Even though this parameter makes the user input for the model a little more complex, it also provides additional context for the implemented large language model, which in turn produces causal evaluation of higher quality.

After the implementation, the proposed model is verified on two datasets—one well known dataset from the area of sensor data collection, which is conventionally used in benchmarking of decision-making methods (labeled as *Sensor* dataset), and one much less known dataset from the domain of graph theory describing properties of graphical structures (labeled as *Graph property* dataset). The datasets were selected for their dissimilarity in areas of expertise and the general familiarity of literature with the datasets. These dissimilarities are crucial to fair verification of causal evaluation done by the selected large language model and also provide a more general idea of the properties of the proposed causal evaluation method.

Using these datasets, the proposed method is experimentally evaluated from three points of view—walkthrough and analysis of visualization component of the model based on Visual Data Analysis and Reasoning (VDAR) evaluation, evaluation of properties of the selected large language model in the context of causal analysis, and comparative analysis of the proposed approach with other, conventionally utilized approaches.

4.1 Evaluation of the Semantical Causal Analysis Outputs

The first of the evaluated aspects is the Causal heatmap visualization conducted using the Visual Data Analysis and Reasoning approach, which focuses on a walkthrough of visual analysis outputs and assessment of visualization's ability to support real, meaningful data reasoning and knowledge discovery. Based on this concept, Figs. 4 and 5 detail the final Causal heatmap visualization for the *Sensor* dataset in the task of identification of causality and the task of identification of new relationships, respectively.

From these visualizations, it is clear that the relationships of the attributes measured in the dataset are well-known and present in the information available to the implemented large language model. This is evident from the fact that the ω values in the causality identification task are all high (the sum of ω values for each of the attributes is 5, meaning each pair of attributes in possible relationships is evaluated as *well known* by the causal evaluator). On the other hand, the familiarity of relationships is confirmed in the task of identification of new relationships in the dataset, where all of the correlation coefficient values were pushed close to 0—caused by the fact that all of the relationships are known. The summarization of the specific overall ω values for individual attributes of the *Sensor* dataset is detailed in Table 1.

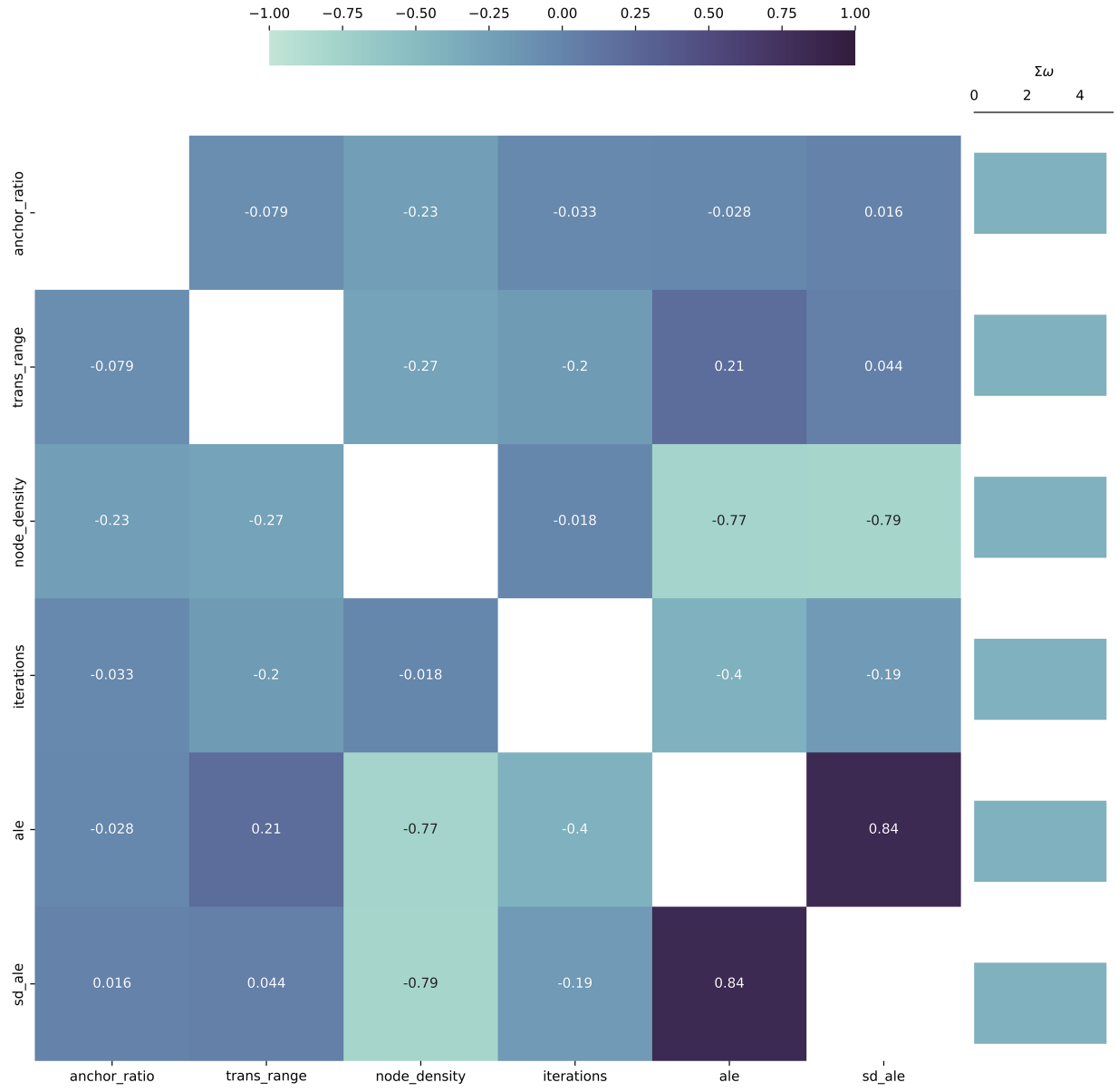

Figure 4: Causal heatmap visualization for sensor dataset with the task of identification of causality.

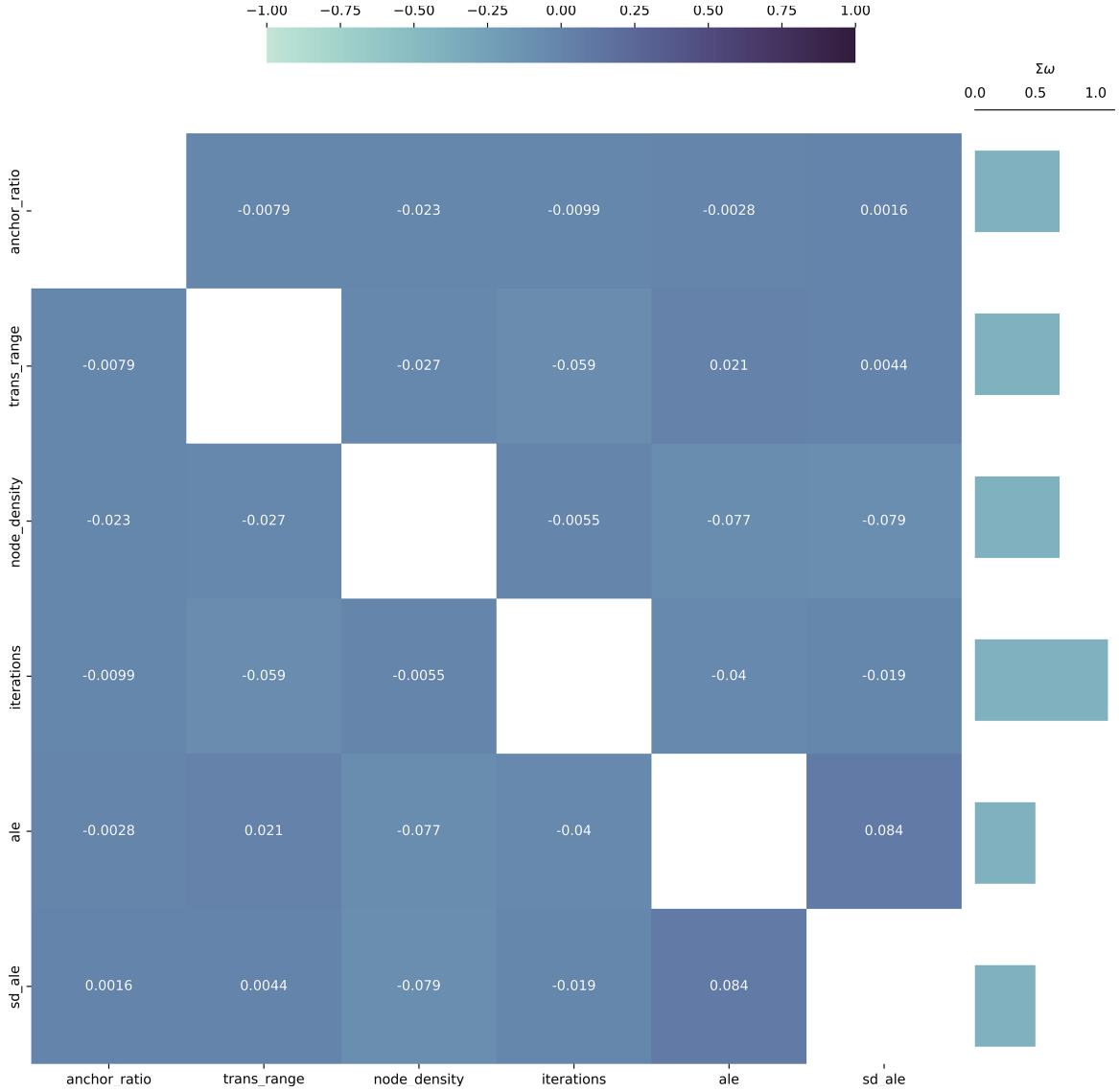

Figure 5: Causal heatmap visualization for sensor dataset with the task of identification of new relationships.

Table 1: Sums of ω values for individual attributes of Sensor dataset.

Attribute	$\sum \omega_{new}$	$\sum \omega_{causal}$
ale	0.5	5
anchor_ratio	0.7	5
iterations	1.1	5
node_density	0.7	5
sd_ale	0.5	5
trans_range	0.7	5

The visualization of Causal heatmaps for both of the considered relationship identification tasks of the *Graph property* dataset is presented in Figs. 6 and 7. As opposed to the *Sensor* dataset, in these cases, the familiarity of available literature with various relationships between graph properties varies significantly, which is reflected in the visualizations. As can be seen, both tasks produce diverse values of $\sum \omega$, which are presented in detail in Table 2.

In both presented dataset studies, a clear pattern is evident—the two evaluated tasks behave in a complementary way for extreme values of ω. Well-known relationships receive high ω scores in the causality-identification setting and, quite naturally, low scores in the new-relationship identification task, while unfamiliar relationships show the opposite trend. This consistency indicates that the large language model correctly distinguishes between familiar and unfamiliar real-world relationships.

Figure 6: Causal heatmap visualization for graph property dataset with the task of identification of causality.

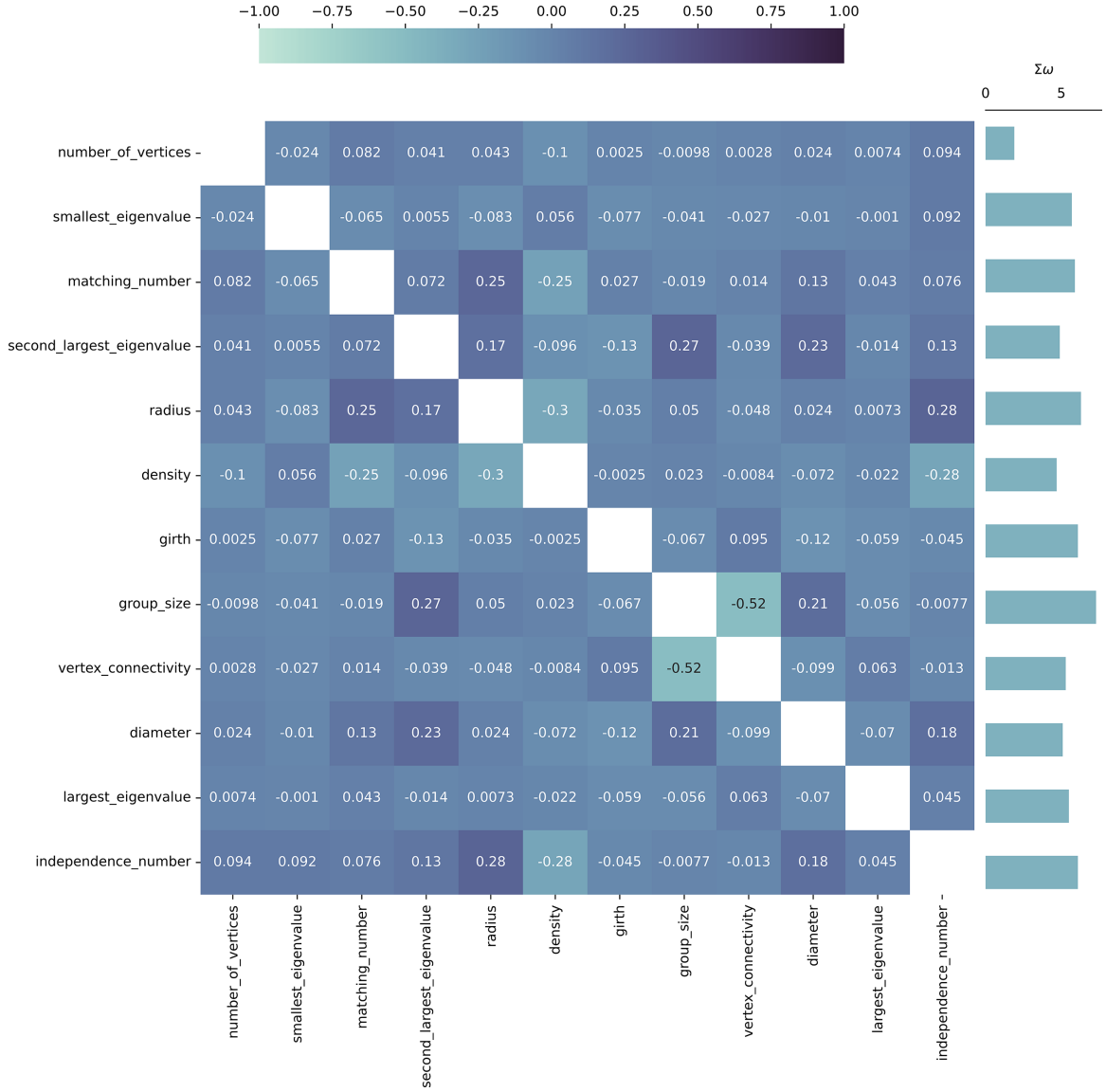

Figure 7: Causal heatmap visualization for graph property dataset with the task of identification of new relationships.

Table 2: Sums of ω values for individual attributes of Graph property dataset.

Attribute	$\sum \omega_{new}$	$\sum \omega_{causal}$
density	4.7	5.4
diameter	5.1	6.5
girth	6.1	5.5
group_size	7.3	2.3
independence_number	6.1	6.3
largest_eigenvalue	5.5	7.2
matching_number	5.9	4.9
number_of_vertices	1.9	9.8

(Continued)

Table 2 (continued)

Attribute	$\sum \omega_{new}$	$\sum \omega_{causal}$
radius	6.3	4.9
second_largest_eigenvalue	4.9	6.9
smallest_eigenvalue	5.7	5.4
vertex_connectivity	5.3	6.1

Qualitatively, the Causal Heatmap visualization demonstrates its usefulness, since relationships of interest are clearly highlighted (using higher coefficient values), while irrelevant relationships are effectively suppressed toward zero, supporting efficient semantic causal analysis via the visualization model.

4.2 Evaluation of the Large Language Model Use as the Semantic Causal Evaluator

The second aspect evaluated in this section of the presented work focuses on the use of large language models as semantic causal evaluators. Since the language model used in the proposed system evaluates the familiarity of relationships present in data as described above, it is critical for these evaluations (generated ω values) to be as consistent as possible. Even though this property is natural from the point of view of the task presented in this study, it is unwanted in most cases which utilize large language models to secure natural language characteristics [14]. Hence, tuning of the input parameters for the large language model is needed—specifically the parameter of temperature $temp \in [0, 2]$ is of interest, since this parameter controls the degree of randomness in token selection [13].

For the purpose of the analysis of generated ω value stability, the model was prompted five times for each of the dataset–type of task combinations, each with three settings for the $temp$ parameter, specifically $temp \in \{0, 1, 2\}$. For each of these experiments the overall sum of ω values of individual attributes ($\sum \omega$), rank of the overall $\sum \omega$ of individual attributes order lowered to highest value ($rank(\sum \omega)$), standard deviation for ω values of attributes ($\sigma(\sum \omega)$), and standard deviation of ranking of attributes ($\sigma(rank)$) was measured. The results of this experimental evaluation are presented in Table 3 for the combination of the *Sensor* dataset and causality identification task, and in Table 4 for the *Graph property* dataset and new relationship identification task.

The results of the experiments conducted on the *Sensor* dataset in the context of the causality identification task clearly point to the high stability of the model when the value of the parameter of $temp = 0$. In this case, both $\sigma(\sum \omega)$ and $\sigma(rank)$ are equal to 0, which is consistent with the definition of the temperature parameter, where the higher the value is, the higher the randomness of the generated responses of the model should be. However, this behaviour is not observed for $temp = 1$ and $temp = 2$, where there are no significant differences between the two—for $temp = 1$ the mean of $\sigma(\sum \omega) = 0.412$ and $\sigma(rank) = 0.982$, while for $temp = 2$ the mean of $\sigma(\sum \omega) = 0.356$ and $\sigma(rank) = 0.997$.

When considering the *Graph property* dataset in the problem of identification of new relationships in data, the behaviour of the tuning is similar for ranking values but differs slightly for ω values themselves. Analysing mean values of ranking, the $\sigma(rank) = 0.286$ for $temp = 0$, and 1.081 and 1.254 for $temp = 1$ and $temp = 2$, respectively, confirming the previous observation. However, the $\sigma(\sum \omega)$ values point to different results – $temp = 0$ produces mean $\sigma(\sum \omega) = 0.563$, $temp = 1$ produces mean $\sigma(\sum \omega) = 0.896$, and $temp = 2$ produces mean $\sigma(\sum \omega) = 0.56$ making the temperature of 0 and 2 more similar to each other, than in the previous case.

Table 3: Stability of ω value for *Sensor* dataset in causality identification task.

temp = 0	$\sum \omega_1$	**rank**$(\sum \omega_1)$	$\sum \omega_2$	**rank**$(\sum \omega_2)$	$\sum \omega_3$	**rank**$(\sum \omega_3)$	$\sum \omega_4$	**rank**$(\sum \omega_4)$	$\sum \omega_5$	**rank**$(\sum \omega_5)$	$\sigma(\sum \omega)$	$\sigma(rank)$
anchor_ratio	3.3	1	3.3	1	3.3	1	3.3	1	3.3	1	0	0
trans_range	3.9	3	3.9	3	3.9	3	3.9	3	3.9	3	0	0
node_density	3.9	3	3.9	3	3.9	3	3.9	3	3.9	3	0	0
iterations	3.7	2	3.7	2	3.7	2	3.7	2	3.7	2	0	0
ale	5	4	5	4	5	4	5	4	5	4	0	0
sd_ale	5	4	5	4	5	4	5	4	5	4	0	0
temp = 1	$\sum \omega_1$	**rank**$(\sum \omega_1)$	$\sum \omega_2$	**rank**$(\sum \omega_2)$	$\sum \omega_3$	**rank**$(\sum \omega_3)$	$\sum \omega_4$	**rank**$(\sum \omega_4)$	$\sum \omega_5$	**rank**$(\sum \omega_5)$	$\sigma(\sum \omega)$	$\sigma(rank)$
anchor_ratio	5	1	4.4	3	3.3	1	4.4	1	3.9	2	0.569	0.8
trans_range	5	1	4.1	2	3.7	2	4.4	1	3.9	2	0.453	0.49
node_density	5	1	4.7	4	3.7	2	4.4	1	4.4	3	0.432	1.166
iterations	5	1	3.8	1	4.1	3	4.4	1	3.8	1	0.449	0.8
ale	5	1	5	5	4.7	4	5	2	5	4	0.12	1.47
sd_ale	5	1	4.4	3	4.7	4	5	2	3.8	1	0.449	1.166
temp = 2	$\sum \omega_1$	**rank**$(\sum \omega_1)$	$\sum \omega_2$	**rank**$(\sum \omega_2)$	$\sum \omega_3$	**rank**$(\sum \omega_3)$	$\sum \omega_4$	**rank**$(\sum \omega_4)$	$\sum \omega_5$	**rank**$(\sum \omega_5)$	$\sigma(\sum \omega)$	$\sigma(rank)$
anchor_ratio	4.1	3	4.4	2	5	1	5	1	4.7	2	0.35	0.748
trans_range	4.1	3	4.4	2	5	1	5	1	4.4	1	0.36	0.8
node_density	4.4	4	4.7	3	5	1	5	1	4.7	2	0.225	1.166
iterations	3.8	2	4.1	1	5	1	5	1	4.4	1	0.48	0.4
ale	4.7	5	5	4	5	1	5	1	5	3	0.12	1.6
sd_ale	3.5	1	5	4	5	1	5	1	5	3	0.6	1.265

Table 4: Stability of ω value for *Graph preperty* dataset in new relationship identification task.

temp = 0	$\sum \omega_1$	rank($\sum \omega_1$)	$\sum \omega_2$	rank($\sum \omega_2$)	$\sum \omega_3$	rank($\sum \omega_3$)	$\sum \omega_4$	rank($\sum \omega_4$)	$\sum \omega_5$	rank($\sum \omega_5$)	$\sigma(\sum \omega)$	$\sigma(rank)$
number_of_vertices	1.3	1	1.3	1	1.3	1	1.3	1	1.3	1	0	0
smallest_eigenvalue	4.1	5	5.7	5	4.1	5	5.7	5	5.7	5	0.784	0
matching_number	6.2	9	8	10	6.2	9	8	10	8	10	0.882	0.49
second_largest _eigenvalue	3.5	2	3.8	2	3.5	2	3.8	2	3.8	2	0.147	0
radius	6.7	10	7.9	9	6.7	10	7.9	9	7.9	9	0.588	0.49
density	4.6	6	5.5	4	4.6	6	5.5	4	5.5	4	0.441	0.98
girth	5.8	8	6.8	8	5.8	8	6.8	8	6.8	8	0.49	0
group_size	6.8	11	8.9	11	6.8	11	8.9	11	8.9	11	1.029	0
vertex_connectivity	3.9	4	5.9	6	3.9	4	5.9	6	5.9	6	0.98	0.98
diameter	3.7	3	4.8	3	3.7	3	4.8	3	4.8	3	0.54	0
largest_eigenvalue	3.5	2	4.8	3	3.5	2	4.8	3	4.8	3	0.637	0.49
independence _number	5.5	7	6	7	5.5	7	6	7	6	7	0.245	0
temp = 1	$\sum \omega_1$	rank($\sum \omega_1$)	$\sum \omega_2$	rank($\sum \omega_2$)	$\sum \omega_3$	rank($\sum \omega_3$)	$\sum \omega_4$	rank($\sum \omega_4$)	$\sum \omega_5$	rank($\sum \omega_5$)	$\sigma(\sum \omega)$	$\sigma(rank)$
number_of_vertices	1.5	1	2.7	1	1.3	1	2.3	1	1.9	1	0.512	0
smallest_eigenvalue	5.1	9	7	8	5.7	10	5.8	9	5.7	7	0.622	1.02
matching_number	5.3	10	7.3	9	6.5	11	6.6	10	6.1	8	0.656	1.02
second_largest _eigenvalue	3.3	4	5.4	4	4.1	5	4	3	4.9	4	0.734	0.632
radius	4.9	8	7	8	4.7	7	5.6	8	6.3	9	0.86	0.632
density	2.7	2	5.4	4	3.5	3	4.4	5	4.8	3	0.956	1.02
girth	4.1	7	6	5	5.1	8	4.8	7	5.2	5	0.615	1.2
group_size	5.7	11	6.3	7	5.3	9	10.7	11	8.4	11	2.016	1.6
vertex_connectivity	3.1	3	5.4	4	3.7	4	3.6	2	4.5	2	0.806	0.894
diameter	3.5	5	5.2	2	2.9	2	4.5	6	5.5	6	0.989	1.833
largest_eigenvalue	3.5	5	5.3	3	2.9	2	4.4	5	4.9	4	0.885	1.166
independence _number	3.7	6	6.2	6	4.3	6	4.3	4	6.4	10	1.102	1.96

(Continued)

Table 4 (continued)

temp = 2	$\sum \omega_1$	rank($\sum \omega_1$)	$\sum \omega_2$	rank($\sum \omega_2$)	$\sum \omega_3$	rank($\sum \omega_3$)	$\sum \omega_4$	rank($\sum \omega_4$)	$\sum \omega_5$	rank($\sum \omega_5$)	$\sigma(\sum \omega)$	$\sigma(rank)$
number_of_vertices	1.3	1	1.7	1	1.5	1	2.3	2	1.9	1	0.344	0.4
smallest_eigenvalue	5.1	8	5.7	9	4.5	6	4.5	7	4.9	7	0.445	1.02
matching_number	6.3	10	4.7	7	6.1	9	4.7	8	5.3	8	0.676	1.02
second_largest _eigenvalue	3.5	3	3.1	4	3.1	2	3.5	5	3.7	3	0.24	1.02
radius	5.9	9	3.7	5	4.7	7	5.1	9	5.5	9	0.755	1.6
density	3.3	2	2.7	2	3.1	2	3.3	4	3.9	4	0.388	0.98
girth	3.9	5	3.9	6	4.5	2	3.7	6	4.7	6	0.388	1.55
group_size	7.2	11	4.9	8	5.3	8	4.5	7	4.7	6	0.977	1.673
vertex_connectivity	4.3	6	2.9	3	3.5	4	3.3	4	3.7	3	0.463	1.095
diameter	4.4	7	2.9	3	3.1	2	3.3	4	4.7	6	0.728	1.855
largest_eigenvalue	3.7	4	3.1	4	3.3	3	1.9	1	3.3	2	0.612	1.166
independence _number	4.7	8	3.1	4	3.7	5	2.7	3	4.1	5	0.709	1.673

Hence, these results show that temperature significantly influences model stability, with $temp = 0$ yielding consistently low variability, while higher temperatures introduce greater randomness, though not always in a consistent way. For the purposes of ω value generation, the stability of the model should be highest possible, and therefore, the $temp$ of 0 should be used to produce a dependable causal evaluation of the relationships.

4.3 Comparative Analysis

As the last part of the evaluation, a comparative analysis is conducted from two perspectives. Firstly, in the context of causal relationship identification, the proposed approach is compared with the Peter-Clark algorithm [15] and the Greedy Equivalence Search method [16]. Secondly, within the context of large language models, the behaviour of Gemini—used as part of the proposed approach—is compared with that of OpenAI ChatGPT [17] and Anthropic Claude [18], as the currently most popular large language models in the area.

Both the Peter-Clark and Greedy Equivalence Search methods identify several causal relationships in the data. However, in comparison to the proposed method, their identification remains purely binary, indicating only the presence or absence of a relationship without providing any evaluation of its strength. On the other hand, the proposed method not only identifies the causal relationships, but also supplements their evaluation via causality coefficient values, enabling the classification of relationships into multiple categories based on their relative strength.

Figs. 8 and 9 present the results obtained using the standard Peter-Clark algorithm and the Greedy Equivalence Search method, respectively, for the $Sensor$ dataset. Both approaches produce similar results, identifying a few notable causal relationships among the dataset attributes. Specifically, both methods identify the relationships $ale - sd\_ale$, $ale - iterations$, $ale - node\_density$, and $sd\_ale - node\_density$ as causal. Additionally, the Peter-Clark algorithm identifies $node\_density - trans\_range$ as a causal relationship of interest, while the Greedy Equivalence Search method includes $sd_ale - trans\_range$ alongside the relationships common to both approaches.

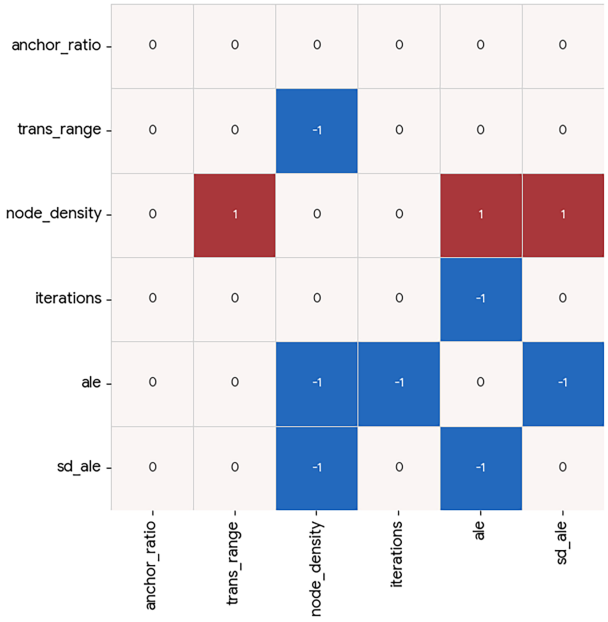

Figure 8: Heatmap of Peter-Clark algorithm for sensor dataset.

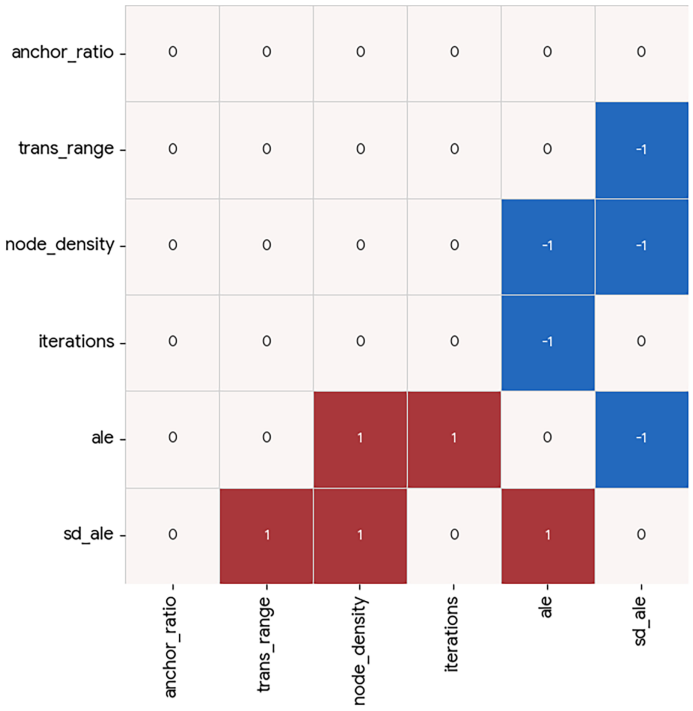

Figure 9: Heatmap of greedy equivalent search approach for sensor dataset.

The method proposed in this study identifies $ale - sd_ale$, $sd_ale - node\_density$, and $ale - node\_density$ as strong relationships of interest, with causality coefficient values of $|caus(attr_A, attr_B)| \approx 0.8$. The relationships $ale - iterations$ and $node\_density - trans\_range$ can be classified as medium to lower-medium strength, with causality coefficients of -0.4 and -0.27, respectively. Finally, the $sd\_ale - trans\_range$ relationship exhibits a low causality coefficient of 0.044, indicating weak causal influence.

For the *Graph Property* dataset, the differences between the results obtained using the proposed approach and those produced by conventional models are more pronounced. Figs. 10 and 11 present the results of the conventional methods, both of which reveal a relatively large set of identified causal relationships. In contrast, the proposed semantic-causal approach identifies a smaller subset of relationships, which can be further categorized according to their inferred strength as follows:

- **Strong causal relationships** include the relationship between $number\_of_vertices$ and $matching\_number$, with a causality coefficient of $caus(number\_of\_vertices, matching\_number) = 0.82$, as well as the relationship between $number\_of\_vertices$ and $independence\_number$, which exhibits a high causality coefficient of 0.94. Additionally, the relationship between $matching\_number$ and $independence\_number$ is classified as strong, with $caus(matching\_number, independence\_number) = 0.76$.
- **Medium-strength relationships** can also be observed in the dataset. These include the relationship between $number\_of\_vertices$ and $radius$, with a causality coefficient of 0.43, and the relationship between $matching\_number$ and $density$, which shows a moderate causal influence with $caus(matching\_number, density) = -0.41$. Furthermore, the $second\_largest\_eigenvalue$ exhibits medium-strength causal relationships with both $vertex\_connectivity$ and $diameter$, characterized by causality coefficients of -0.39 and 0.53, respectively. Finally, the relationship between $density$ and $independence\_number$ can also be classified as medium strength, with a causality coefficient of -0.66.

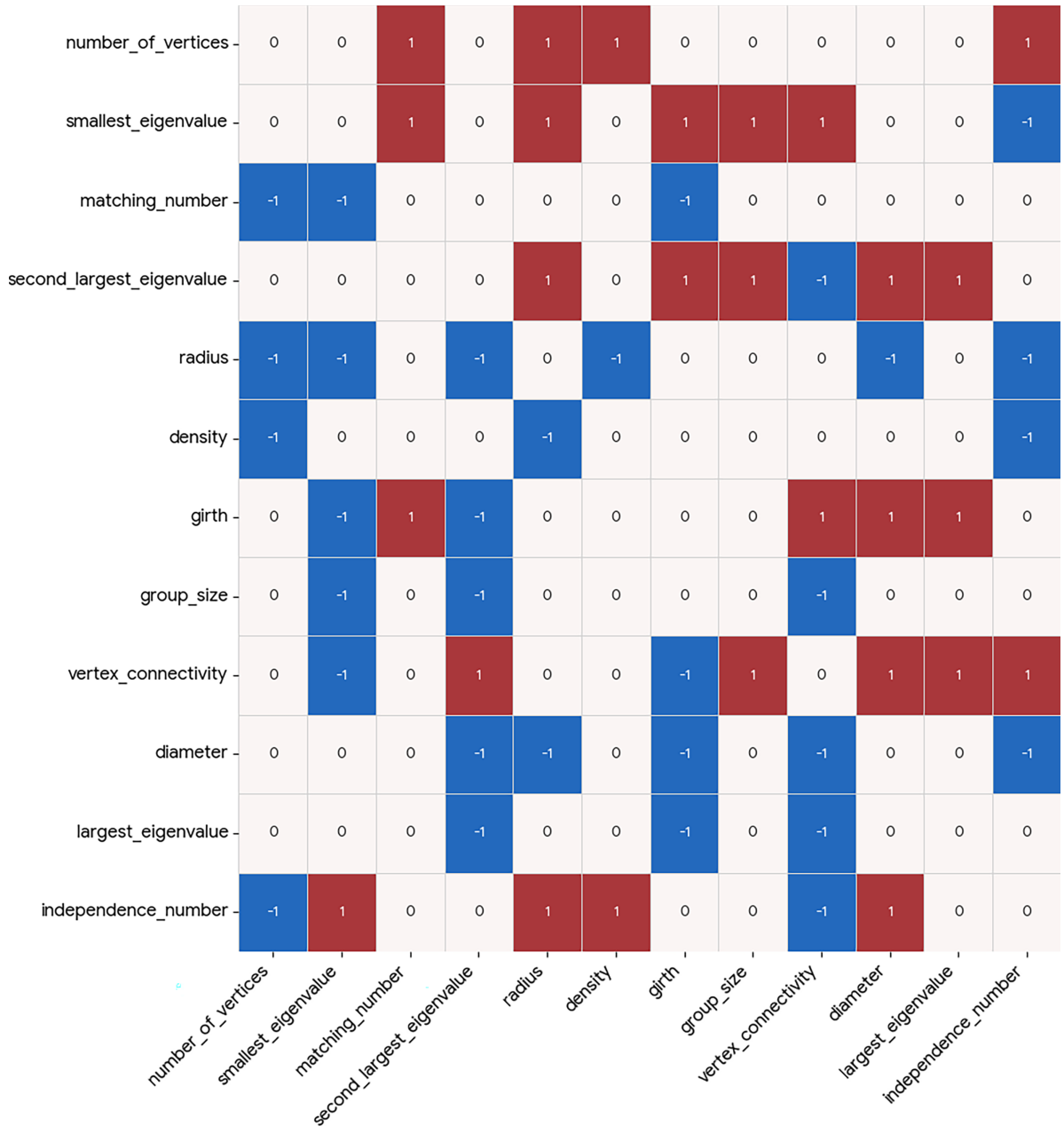

Figure 10: Heatmap of Peter-Clark algorithm for graph property dataset.

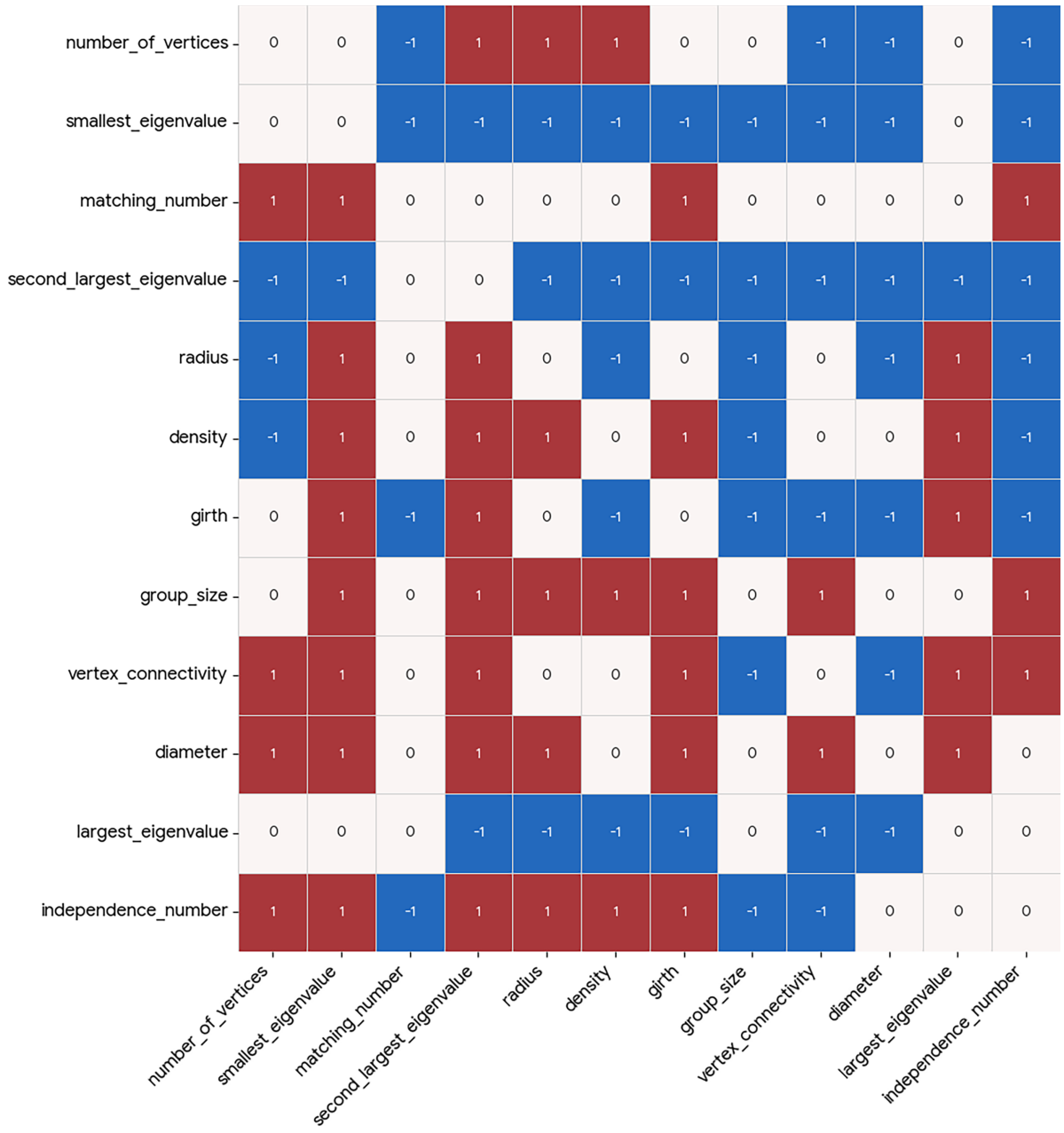

Figure 11: Heatmap of greedy equivalent search approach for graph property dataset.

When comparing the results of the proposed method with those obtained using conventional approaches, it becomes evident that the original proposed method presents more strict strategy to the identification of causal relationships. However, besides the identification of relationships themselves, the proposed method evaluates them quantitatively through the use of causality coefficient values, providing a more nuanced assessment than binary causal discovery methods.

The second aspect of comparative analysis focuses on the utilization of large language models in the evaluatory process. Table 5 presents a comparison of ω values and their corresponding rankings for semantic-causal evaluation of attributes in the *Sensor* dataset obtained using OpenAI ChatGPT and Anthropic Claude. Overall, the results and stability of the evaluated models are largely comparable across all three of the considered large language models. Significant differences between the models are present in terms of output consistency and suitability for automation necessary for the considered task. ChatGPT disregards the specified output formatting in approximately 30% of cases, frequently introducing additional explanatory text or introductory sentences. Such behaviour reduces the feasibility of automated processing, which is a critical requirement of the proposed approach. Furthermore, Claude consistently ignores the formatting constraints defined in the prompt, rendering full automation not possible.

Table 6 presents the corresponding comparison for the *Graph_Property* dataset, where more pronounced differences among the models are observable. In this case, ChatGPT assigns an identical ω value of 0.5—indicating a neutral relationship—to all evaluated attribute pairs. This, combined with additional model outputs, points to the fact, that the model is unable to perform meaningful evaluation under the given prompt and would require additional contextual information. Conversely, Claude produces ω values and rankings that are broadly consistent with those generated by Gemini; however, it continues to disregard the required output formatting, limiting its applicability within an automated evaluation pipeline.

Table 5: ω value for $Sensor$ dataset in causality identification task using alternative large language models.

ChatGPT	$\sum \omega_1$	**rank**($\sum \omega_1$)	$\sum \omega_2$	**rank**($\sum \omega_2$)	$\sum \omega_3$	**rank**($\sum \omega_3$)	$\sum \omega_4$	**rank**($\sum \omega_4$)	$\sum \omega_5$	**rank**($\sum \omega_5$)	$\sigma(\sum \omega)$	$\sigma(rank)$
anchor_ratio	3.9	3	3.1	2	3.2	3	4.2	5	3.6	4	0.41	1.02
trans_range	3.6	2	3.1	2	3.6	4	3.9	4	2.9	2	0.37	0.98
node_density	3.6	2	3.3	3	3.1	2	3	2	2.9	2	0.25	0.4
iterations	2.3	1	2.7	1	2.9	1	2	1	2.5	1	0.31	0
ale	4.3	4	3.8	5	4.4	5	4.4	6	3.9	5	0.26	0.63
sd_ale	3.9	3	3.6	4	3.6	4	3.6	3	3.2	3	0.22	0.49
Claude	$\sum \omega_1$	**rank**($\sum \omega_1$)	$\sum \omega_2$	**rank**($\sum \omega_2$)	$\sum \omega_3$	**rank**($\sum \omega_3$)	$\sum \omega_4$	**rank**($\sum \omega_4$)	$\sum \omega_5$	**rank**($\sum \omega_5$)	$\sigma(\sum \omega)$	$\sigma(rank)$
anchor_ratio	5	4	3.9	3	3.4	2	3.4	2	3.8	3	0.59	0.75
trans_range	4.4	3	3.9	3	3.4	2	3.4	2	3.8	3	0.37	0.49
node_density	4.2	2	3.9	3	3.4	2	3.4	2	3.8	3	0.31	0.49
iterations	4.2	2	3.7	2	3.1	1	3.1	1	3.5	2	0.41	0.49
ale	4.4	3	4.3	4	4.2	4	4.3	3	4.3	4	0.06	0.49
sd_ale	3.9	1	3.3	1	3.5	3	3.4	2	3.3	1	0.22	0.8

Table 6: ω value for *Graph property* dataset in causality identification task using alternative large language models.

ChatGPT	$\sum \omega_1$	**rank**$(\sum \omega_1)$	$\sum \omega_2$	**rank**$(\sum \omega_2)$	$\sum \omega_3$	**rank**$(\sum \omega_3)$	$\sum \omega_4$	**rank**$(\sum \omega_4)$	$\sum \omega_5$	**rank**$(\sum \omega_5)$	$\sigma(\sum \omega)$	$\sigma(rank)$
number_of_vertices	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
smallest_eigenvalue	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
matching_number	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
second_largest_eigenvalue	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
radius	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
density	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
girth	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
group_size	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
vertex_connectivity	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
diameter	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
largest_eigenvalue	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
independence_number	5.5	1	5.5	1	5.5	1	5.5	1	5.5	1	0	0
Claude	$\sum \omega_1$	**rank**$(\sum \omega_1)$	$\sum \omega_2$	**rank**$(\sum \omega_2)$	$\sum \omega_3$	**rank**$(\sum \omega_3)$	$\sum \omega_4$	**rank**$(\sum \omega_4)$	$\sum \omega_5$	**rank**$(\sum \omega_5)$	$\sigma(\sum \omega)$	$\sigma(rank)$
number_of_vertices	3.9	1	2.5	1	3.7	1	5.2	6	3.4	1	0.87	2
smallest_eigenvalue	6.1	9	3.9	7	4.5	5	5.4	8	5.5	9	0.78	1.5
matching_number	5.5	8	4.1	8	5.3	8	5.9	11	5.8	10	0.65	1.27
second_largest_eigenvalue	6.3	10	3.9	7	4.7	6	5.3	7	5.2	7	0.79	1.36
radius	4.7	4	3.7	6	4.5	5	4.7	3	4.6	5	0.38	1.02
density	4.1	2	2.9	3	4.3	4	4.8	4	4.4	4	0.64	0.8
girth	5.3	7	4.5	9	5.1	7	5.1	5	5.9	11	0.45	2.04
group_size	7.7	11	5.9	10	6.1	9	7.7	12	7.7	12	0.84	1.17
vertex_connectivity	4.9	5	3.5	5	4.7	6	5.5	9	5.3	8	0.7	1.63
diameter	4.1	2	3.1	4	4.1	3	4.5	2	4.3	3	0.48	0.75
largest_eigenvalue	4.5	3	2.7	2	3.9	2	4.4	1	4	2	0.64	0.63
independence_number	5.1	6	3.7	6	4.5	5	5.6	10	5.1	6	0.65	1.74

5 Conclusion

Correlation and causality are frequently examined together, as correlations often motivate deeper causal investigation of the situation identified in data. This study introduces a visual analysis method that utilizes large language models to assess the semantic causality of relationships initially identified through correlation analysis. The proposed approach was experimentally evaluated from two perspectives—its visualization design and its large language model-based causal evaluation, where both of the components produced results deemed satisfactory, while several advantages and disadvantages of the method became evident.

The proposed approach offers several advantages, including automatic semantic causality evaluation rooted in real-world familiarity of relationships, clear highlighting of causal or novel relationships while suppressing weak ones, and the ability to highlight attributes where relationships are well known or largely unexplored. The main disadvantages involve a degree of unstable behaviour of the large language model, which is expected to diminish as future large language model generations improve, and the need for proper labelling of attributes to ensure reliable assessment of the real-world plausibility of relationships.

In future work conducted within the studied area, the visualization of the direct and indirect semantic causality based on large language model causal evaluation needs to be explored via graphical models—two of such causal structures are obvious as extensions of the presented work—causal graphs and causal chains. Secondly, the 5-tier classification of the familiarity of a relationship used in this study can be further replaced with 7-, 9-, or more tiered approaches; the influence on semantic causal evaluation of which needs to be examined.

Acknowledgement: Not applicable.

Funding Statement: The research presented in the study was supported by University Grant Agency of Matej Bel University in Banská Bystrica project number UGA-14-PDS-2025.

Availability of Data and Materials: The code for the proposed causality visualization method and Sensor dataset used in the presented experiments are openly available at: https://github.com/AdamDudasUMB/CausalEvaluation. For additional information, contact the author at adam.dudas@umb.sk.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Jiang Y, Feng XJ, Hue WG, Wang P. An intelligent causality analysis system for aviation safety based on nonaxiomatic logic graphs. Intell Comput. 2025;4(1):180. doi:10.34133/icomputing.0180.
2. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The rise and potential of large language model based agents: a survey. Sci China Inf Sci. 2025;68(2):121101. doi:10.1007/s11432-024-4222-0.
3. Hsu HP. From programming to prompting: developing computational thinking through large language model-based generative artificial intelligence. TechTrends. 2025;59(3):482–506.
4. Liu PF, Ren YM, Tao J, Ren ZX. GIT-Mol: a multi-modal large language model for molecular science with graph, image, and text. Comput Biol Med. 2024;171:1–11.
5. Zhao TQ, Yogaratham A, Yue M. A large language model for determining partial tripping of distributed energy resources. IEEE Trans Smart Grid. 2025;16(1):437–40. doi:10.1109/tsg.2024.3453649.
6. Quan H, Ni W, Zhang T, Ye X, Xie Z, Wang S, et al. Large language model agents for radio map generation and wireless network planning. IEEE Netw Lett. 2025;7(3):166–70. doi:10.1109/lnet.2025.3539829.
7. Zhan Y, Xiong Z, Yuan Y. Unifying remote sensing vision-language tasks via instruction tuning with large language model. ISPRS J Photogramm Remote Sens. 2025;221(8):64–77. doi:10.1016/j.isprsjprs.2025.01.020.

8. Dudáš A, Vagač M. Diagnostic analysis approach to correlation maps through large language models. In: Proceedings of the 2024 IEEE 17th International Scientific Conference on Informatics (Informatics); 2024 Nov 13–15; Poprad, Slovakia. p. 62–8.
9. Vagač M, Dudáš A. Web application for large language model-based diagnostic analysis of correlation maps. IPSI BGD Trans Internet Res. 2025;21(1):62–75. doi:10.58245/ipsi.tir.2502.07.
10. Iantovics LB. Avoiding mistakes in bivariate linear regression and correlation analysis, in rigorous research. Acta Polytech Hung. 2024;21(6):33–52. doi:10.12700/aph.21.6.2024.6.2.
11. Yu H, Hutson AD. A robust Spearman correlation coefficient permutation test. Commun Stat Theor Meth. 2024;53(6):2141–53. doi:10.1080/03610926.2022.2121144.
12. Annoye H, Beretta A, Heuchenne C. Statistical matching using autoencoders-canonical correlation analysis, kernel canonical correlation analysis and multi-output multilayer perceptron. Knowl Based Syst. 2025;330(B):114626. doi:10.1016/j.knosys.2025.114626.
13. Google. Gemini API documentation; 2025 Sep 8. [cited 2026 Feb 1]. Available from: https://ai.google.dev/api.
14. Tao Y, Yang R, Wen Y, Zhong Y, Jiao K, Gu X. LLM-KE: an ontology-aware LLM methodology for military domain knowledge extraction. Comput Mater Contin. 2026;86(1):1–17. doi:10.32604/cmc.2025.068670.
15. Biswas R, Mukherjee S. Consistent causal inference from time series with PC algorithm and its time-aware extension. Stat Comput. 2024;34(1):14. doi:10.1007/s11222-023-10330-3.
16. Liu X, Feng Q, Yang Z, Wu S, Gao X, Yang Y, et al. An improved greedy equivalent search method based on relative entropy. Sci Rep. 2025;15(1):37250. doi:10.1038/s41598-025-21219-8.
17. OpenAI. GPT API documentation. [cited 2026 Feb 1]. Available from: https://platform.openai.com/docs.
18. Anthropic. Claude API documentation. [cited 2026 Feb 1]. Available from: https://platform.claude.com/docs.