

Article

Large Language Models as Semantic Evaluators of Embedded Correlation Substructures

Adam Dudáš *  and Peter Babic

Department of Computer Science, Faculty of Natural Sciences, Matej Bel University, 974 01 Banská Bystrica, Slovakia; peter.babic@student.umb.sk

* Correspondence: adam.dudas@umb.sk

Abstract

Graphical methods of correlation analysis, such as correlation n -ptychs or hotspots, focus on the identification of the strength and direction of functional relationships between sets of attributes in multidimensional datasets. Since these correlation structures only take into account values of the attributes, situations arise when the relationship is coincidental, meaning that there is no real-world causality between the values of the observed attributes but these values still exhibit significant correlation. This problem of correlation analysis as a whole motivates the need for semantic evaluation of significant relationships identified using its methods—a task that could potentially be time- and resource-intensive when conducted manually. However, modern results in the large language model area provide tools for the automatization of such tasks. Hence, this work focuses on the design and implementation of a novel large language model-based method for semantic evaluation of correlation structures embedded in a correlation graph, specifically correlation n -ptychs for $n \in \{3, 4, 5\}$ and correlation hotspots. In the method, the large language model is automatically prompted to assess the semantic nature of relationships in the set of correlation substructures of the dataset, identify their real-world relevance, and visualize the result in the form of a Semantic evaluation card. The proposed approach is evaluated using two benchmarking datasets focusing on the visualization method used in the model, large language model interaction with the correlation substructures, and comparative analysis with previously used tools in the area.

Keywords: correlation structures; correlation hotspots; correlation n -ptychs; large language models; visualization; semantic evaluation



Academic Editor: Takayuki Hibi

Received: 17 April 2026

Revised: 29 May 2026

Accepted: 9 June 2026

Published: 11 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Datasets of high dimensionality present their specific problems for any type of data analysis by their structure alone. Since such datasets are characterized by their high number of attributes, even simple tasks such as dataset description by aggregation, correlation analysis, or data visualization can be frequently problematic to conduct. This motivated the development of graphical models for visualization of direct and indirect relationships in the data based on correlation analysis, called correlation structures [1].

All correlation structures are composed of a set of vertices, which represent individual attributes of the dataset, and a set of edges between pairs of these attributes, which are weighted by the value of the correlation coefficient measured between the given attribute pairs. For the purposes of deeper correlation analysis of a dataset, two correlation substructures are utilized:

- Correlation n -ptychs—complete K_n subgraphs (cliques) of the main correlation graph of n attributes [2].
- Correlation hotspots—complete bipartite $K_{1,h}$ graphs of the main correlation graph [3].

However, as with any correlation analysis method, correlation structures can identify coincidental relationships, which are not representative of the real world [4]. This motivates the inclusion of the study of the semantics of these relationships for a greater understanding of the examined data. If done manually, such an analysis of relationship semantics would be largely resource-intensive and frequently impossible for human experts to do in a practical time. With the fast development of large language models, which are well documented in situations of replacing human experts in simple problem-solving [5,6], the process of semantic evaluation can be automated. Utilizing such a language model, the relationships identified in data by correlation analysis can be easily classified into semantically supported, likely causal relationships where the real-world aspects of the problem can be observed and studied, and non-causal, which are not known in the knowledge corpus available to the language model.

1.1. Objective of the Work

The problems mentioned above motivate the development of an exploration approach focused on the semantic evaluation of relationships in correlation substructures. Therefore, the main objective of this work can be summarized as the design and implementation of an evaluator of correlation n -ptychs and hotspots that utilizes large language models to assess the semantic quality of the relationships in these substructures. The model focuses on sets of three or more attributes of the studied dataset, between which significant correlation coefficient values are measured, and classifies them into two main categories based on their representativeness of the real-world relationships of the two parameters. Such an evaluation of the relationships is complemented by the visualization of the specific correlation substructures and linguistic description of the decision generated by a large language model.

The objectives of the presented study can be summarized as follows:

- The design and implementation of the proposed large language model-based semantic evaluation method of correlation substructures.
- The design and implementation of the visualization approach for presentation of the results of such a semantic evaluation of relationships via Semantic evaluation cards.
- A qualitative, quantitative, and comparative evaluation of the proposed model with the use of two openly available benchmarking datasets.

The rest of the work presented in the context of this study is structured into three main sections. After the part of the text devoted to the brief overview of works from the studied area, which is presented at the end of the current section, Section 2 focuses on the description and semantic evaluation of correlation substructures, and presents the proposed method for its conducting. Key aspects of this method—(MDPI: We changed minus sign to emdash. Please confirm. The same below.)the visualization approach, large language model use, and comparison to similar approaches—are then evaluated in Section 3. Lastly, Section 4 concludes the work and specifies the possible future work areas in the studied domain.

1.2. Related Works

Since the method proposed in this study utilizes large language models in the context of diagnostic analysis of graphical models from the data science area, which is conventionally done by human experts, a brief overview of related work utilizing these language models is presented.

Numerous studies point to the relevance of the use of large language models as human expert surrogates in various domains. As an example, the research presented in [7] focuses on the performance of large language models (specifically, ChatGPT 4) in the analysis of political texts, while comparing its properties to human experts from the area. The results of the study highlight high accuracy of the language models in the task (up to 95%) with a slight decline in performance for longer texts and non-English content. The authors conclude that the quality of text analysis is comparable between human experts and large language models, while the latter offers additional advantages, such as a lower required time and the possibility of automation.

Similarly, [8] utilizes large language models in the context of predictive tasks from the neuroscience domain. The authors show that, using specialized tools, large language models outperform human experts significantly while behaving similarly to humans regarding decision-making confidence, which correlates with accuracy. Additionally, the authors propose an approach of generalization of the used methods to other knowledge-intensive domains.

In the area of data science, the work [9] studies the impact of large language models on solving several typical analytical problems, such as data exploration, visualization, association rule mining, or classification. The authors highlight the potential of large language models in the automation of some of the stages of data analysis, which in turn allows for acceleration in analytical processes and decision-making. Interestingly, the authors also point to limitations of such approaches, which can arise in complex scenarios.

Our previous work in the area of large language model utilization in the context of correlation analysis conducted over multidimensional datasets proposed two visual methods for a more detailed view of the data.

In [10], the focus is on the design and implementation of a model for diagnostic data analysis, which linguistically describes correlated pairs of attributes using a large language model. The results of this analysis are presented in the context of diagnostic sheet visualization, which combines an interactive correlation heatmap and large language model-generated descriptions of interesting attribute pairs of the dataset (see Figure 1).

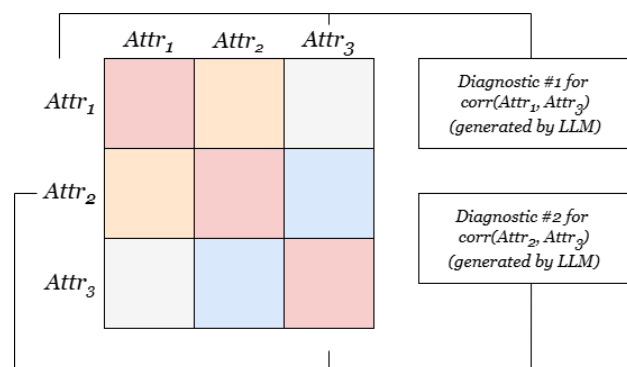


Figure 1. Schematic visualization of the diagnostic sheet approach presented in [10].

A more recent advance in the selected area is presented in the work [11], where large language models are used as semantic causality evaluators of correlation matrices presented in the form of causal heatmaps (Figure 2). The approach combines two aspects of semantic analysis of relationships in a dataset: firstly, each correlation coefficient value is weighted using the ω -value generated by the large language model, which signifies the level of real-world familiarity of the relationship between the correlated attributes; secondly, the results are visualized using a heatmap supplemented by $\sum(\omega)$ for each attribute. This metric summarizes the number of real-world relationships connected to the given attribute.

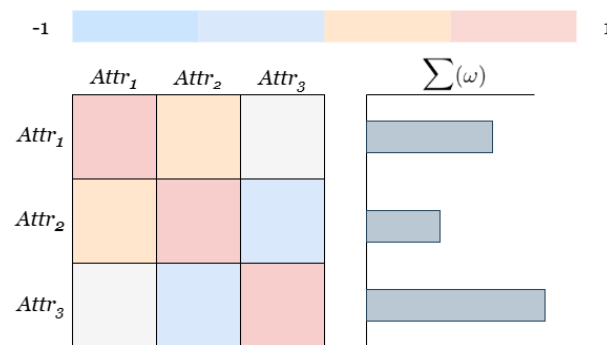


Figure 2. Schematic visualization of the causal heatmap approach presented in [11].

In the presented study, two of the previously proposed methods are supplemented with the Semantic evaluation cards approach, focused on the diagnostic analysis of correlated subsets of attributes utilizing large language models.

2. Semantic Evaluation of Correlation Substructures

The research presented in this work is based on the correlation analysis principles, in which correlation determines the strength and direction of the functional relationship between a pair of quantitative attributes $Attr_A$ and $Attr_B$. In the context of data analysis, such a relationship determined in correlation analysis is interpreted as the prediction potential stored between the two attributes of the studied dataset, and is measured via correlation coefficients $corr(Attr_A, Attr_B) \in [-1, 1]$ of various types (most notably the Pearson correlation coefficient and Spearman rank correlation coefficient) [12].

Regardless of its type, the correlation coefficient always measures prediction potential between a pair of attributes, which needs to be extended if studying datasets of high dimensionality, as considered in this study. Often, correlation matrices are used to summarize pairwise relationships in the dataset as shown [13]:

$$\mathbf{C} = \begin{bmatrix} corr(Attr_A, Attr_A) & corr(Attr_A, Attr_B) & \cdots & corr(Attr_A, Attr_Z) \\ corr(Attr_B, Attr_A) & corr(Attr_B, Attr_B) & \cdots & corr(Attr_B, Attr_Z) \\ \vdots & \vdots & \ddots & \vdots \\ corr(Attr_Z, Attr_A) & corr(Attr_Z, Attr_B) & \cdots & corr(Attr_Z, Attr_Z) \end{bmatrix} \quad (1)$$

As seen, the correlation matrix is cross-indexed by attributes of the dataset, while each of its elements contains the value of the correlation coefficient measured between the respective indexing attributes. In such a correlation matrix, analysts can easily identify significant direct relationships in data, while indirect ones are not immediately clear. This motivated the correlation structure approaches, which use the correlation matrix as the adjacency matrix of a complete, weighted, undirected graph, called a correlation graph.

Complete correlation graph G is graphical model consisting of sets V and E , where V is a set of vertices, each representing one of the quantitative attributes of the studied dataset, and E denotes a set of edges, which connect vertices of this graph, and which are weighted using the correlation coefficient measured between the given pair of attributes (starting and ending vertices of the edge) [1]. As seen in the left subfigure of Figure 3, even low-dimensional complete correlation graphs can be hard to read and visualize intuitively, since the graph itself is not planar in the case when the dataset contains five or more attributes.

Since not all of the correlation coefficient values measured in the dataset—and visualized in the correlation graph—are of significance to the analysis of data, the concept of pruning of the correlation graph has been introduced. By pruning of the graph, the analyst applies masking on the correlation matrix of the dataset, which acts as a filter of correlation

coefficient values, and therefore edges in the graph. Several correlation graph pruning approaches have been introduced, notably the extended σ -mask pruning of correlation matrix $\mathbb{C}$ [1]:

$$\sigma(\mathbb{C}) = \frac{\mu(\mathbb{C}) + \max(\mathbb{C})}{2} + \alpha \quad (2)$$

where $\mu(\mathbb{C})$ denotes the mean of absolute correlation coefficient values in the matrix not taking into account the main diagonal, $\max(\mathbb{C})$ denotes the maximal value of absolute correlation coefficients in the matrix not taking into account the main diagonal, and α is the user-defined mask strictness parameter, where $\alpha \in [0, 0.3]$.

Then, the σ -mask pruning $\forall \text{corr}(\text{Attr}_A, \text{Attr}_B) \in \mathbb{C}$ is conducted as

$$\text{corr}(\text{Attr}_A, \text{Attr}_B) = \begin{cases} \text{significant}, & \text{if } |\text{corr}(\text{Attr}_A, \text{Attr}_B)| \geq \sigma, \\ \text{insignificant}, & \text{otherwise} \end{cases} \quad (3)$$

while *significant* correlation coefficient values are left in the pruned space and *insignificant* values are filtered out. This operation creates the pruned correlation graph, which represents the set of significant relationships in the dataset and their interaction (see Figure 3). Such a pruned correlation graph is then used as a basis for the identification of correlation substructures relevant for this study.

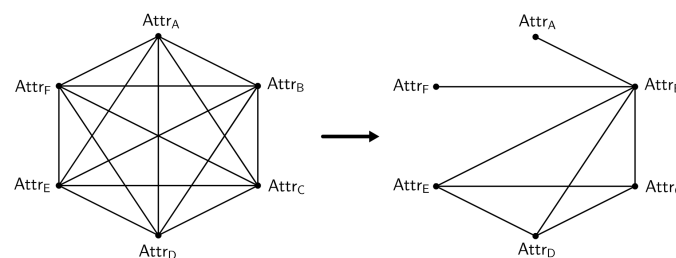


Figure 3. Example of complete correlation graph (left) and its pruning (right) for a dataset of six attributes.

2.1. Correlation n -Ptychs and Hotspots

There are several embedded graphical substructures, which can be identified in the pruned correlation graph G , such as correlation chains, correlation cycles, or in the case of this work, correlation n -ptychs and correlation hotspots.

A complete subgraph of G with $n \in \{3, 4, 5\}$ vertices is referred to as an n -vertex subgraph. Such an n -ptych represents a set of n attributes, in which the values of each of the attributes directly influence the values of all other attributes in the set. Hence, the values of the subset of attributes are tightly connected, which is a property relevant for tasks such as feature selection—these related attributes can be either reduced to one attribute since their values are redundant from the point of view of new information supplemented to the machine learning model, or, vice-versa, their values can be used to compute missing values for attributes in the set, since the relationships in it are strong [2].

On the other hand, a correlation hotspot is a complete bipartite graph $K_{1,h}$ where one central vertex is connected to a set of h leaves that are not interconnected themselves. Therefore, the correlation hotspot is a star schema, where a single central attribute correlates with multiple leaf attributes that, however, do not correlate directly with one another. This substructure may suggest the existence of a controlling factor or common cause that influences the other attributes within the topology. The central attribute of a hotspot can be considered an attribute that is significant in terms of the dataset's predictive potential, since its values significantly influence the values of a set of other attributes that do not exhibit any measurable relationship among themselves [3].

As seen in Figure 4, both correlation n -ptychs and hotspots can be identified in the pruned correlation graph using simple path- and subgraph-finding algorithms, such as the Bron–Kerbosch algorithm [14] or maximal biclique enumeration algorithm [15].

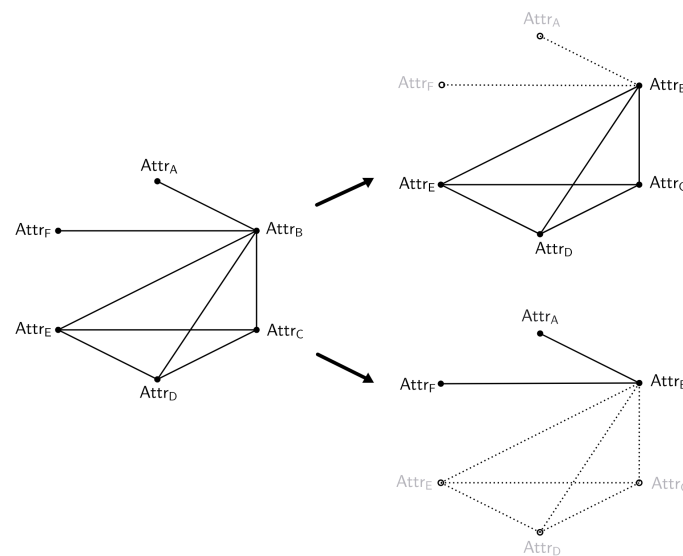


Figure 4. Example of correlation n -ptych of size 4 (**top**) and correlation hotspot of size 2 (**bottom**) embedded in a pruned correlation graph of six attributes.

2.2. Semantic Evaluation of Correlation Substructures Using Large Language Models

Based on the identification of correlation substructures described by a set of attributes from the studied dataset, a semantic evaluation of the relationships within these sets needs to be performed in order to identify relationships that are semantically supported by domain knowledge and coincidental ones. As stated above, in this work, an approach to this task that combines large language model-based semantic evaluation with interactive visualization of correlation substructures, called Semantic evaluation cards, is introduced. The proposed model consists of several steps, illustrated in the schema in Figure 5, with the input of the model being the studied dataset along with brief descriptions of its attributes, which are used to guide the later phases of the process.

The first step of the proposed process consists of the construction of the correlation graph. The model performs correlation analysis and summarizes the resulting coefficients into a correlation matrix. This matrix is then pruned using extended σ -masking, producing a pruned correlation graph. This step involves internal parameters for the process, namely the type of correlation coefficient (with Pearson correlation coefficient used as a default value) and the α value (set to 0.1 by default).

Next, the model identifies correlation substructures within the graph. Specifically, the n -ptychs of size $n \in \{3, 4, 5\}$ and all hotspots embedded in the graph are identified. These identified substructures are forwarded to both the diagnostic analysis stage and to the final output of the process.

Besides the set of correlation substructures embedded in the pruned correlation graph of the dataset, the diagnostic analysis step of the proposed process also expects the description of attributes in the dataset. Then each substructure is passed to the large language model, which offers a semantic analysis of the relationships in the substructure from the point of view of real-world familiarity with the domain and the given attributes. The prompting strategy used involves separate, tailored prompts for n -ptychs and hotspots. The n -ptych prompt consists of the following:

Context: The following attributes exhibit a special relationship: every attribute in this group is strongly and directly correlated with every other attribute.

Attributes: {format_attributes(subgraph_nodes, descriptions)}

Question: Is there a clear underlying physical phenomenon or shared system that explains this tight, mutual correlation?

Instructions: You MUST start your answer with 'Yes,' followed by a concise, one-sentence hypothesis, or 'No,' if a direct link is unlikely.

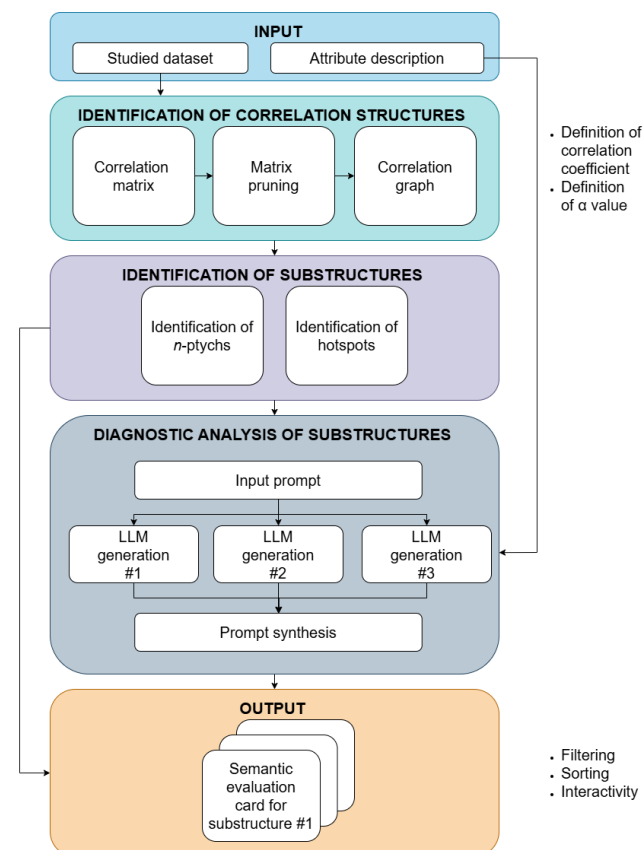


Figure 5. Flowchart of the proposed semantic evaluation process.

For the purposes of the correlation hotspot prompting, the following wording is used:

Context: The following attributes exhibit a 'hub-and-spoke' relationship. The central attribute '{central_node}' is strongly correlated with each of the 'leaf' attributes: {' , ' .join(leaf_nodes)}. However, the leaf attributes are not strongly correlated with each other.

Attributes: {format_attributes(subgraph_nodes, descriptions)}

Question: Is there a clear causal or influential relationship here, with '{central_node}' acting as a driver?

Instructions: You MUST start your answer with 'Yes,' and then explain the role of the central node '{central_node}', or 'No,' if the relationship is not clear.

In the prompting, an ensemble-like, three-fold approach to prompt-response generation, which improves the robustness and consistency of the large language model outputs, is implemented [16]. Specifically, in the presented method, each prompt is sent to the model three times in parallel, and the resulting—partial—responses are subsequently synthesized into a single answer. However, to ensure full transparency of the process, the final output presents the analyst with all four responses (the final synthesized response and three partial ones).

The output of the model consists of both the diagnostic analysis results and the graphical visualization of the identified correlation substructures, which are presented in the form of a Semantic evaluation card (see Figure 6). The final output of the method contains a collection of such cards within an interactive environment. This setup enables users to filter and sort the results, as well as to interact with both the graphical elements of the visualization and the cards themselves.

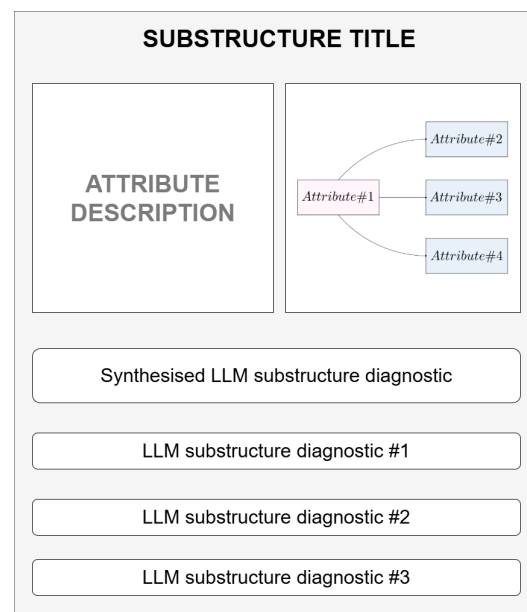


Figure 6. Schematic visualization of the detailed Semantic evaluation card of a correlation substructure.

3. Evaluation of the Semantic Evaluation Card Environment

The proposed semantic evaluation of correlation substructure-based relationships identified in multidimensional datasets is implemented in *Python* using *pandas* 3.0.1 and *numpy* 2.4.0 packages for the basic operation with loaded data, *networkx* 3.6.1 and *matplotlib* 3.10.0 for visualization purposes, and *llama-cpp-python* 0.3.20 for the implementation of the *Mistral-7B-Instruct-v0.2* language model used in the semantic analysis of the data [17]. The use of this model was motivated by its low hardware requirements, strong instruction-following capabilities in long contexts, and fast inference when compared to similar size models [18,19].

For evaluation purposes, the implementation of the proposed method is evaluated using two openly available benchmarking datasets from two different domains:

- Appliances Energy Prediction dataset (*AppEn*)—the first of the used datasets contains experimental data measurements of the use of energy by appliances in a low-energy building in Belgium. The dataset contains 28 attributes focused on the energy consumption, temperature, and humidity in various rooms of the building, or external environmental measurements [20]. The dataset contains the following attributes [21]:
 - Internal temperature numbered by the rooms of the building—*T1* to *T9*.
 - Internal relative humidity numbered by the rooms of the building—*RH1* to *RH9*.

- Timestamp of the data measurement stored in *date*.
- Energy use of appliances in the house stored in *appliances*.
- Energy use of light fixtures in the house stored in *lights*.
- External measurements of temperature (T_{out}), relative humidity (RH_{out}), atmospheric pressure ($Press_{mm_hg}$), speed of wind ($Windspeed$), visibility with naked eye $Visibility$, and dewpoint temperature $T_{dewpoint}$ from Chievres weather station located close to the building.
- Two control random noise attributes *rv1* and *rv2*.
- Superconductivity Data dataset (*SuperCond*)—the second dataset used in the evaluation process focuses on chemical and physical properties of superconductors [22]. Compared to the *AppEn* dataset, *SuperCond* contains a significantly higher number of attributes, specifically 81, which focus on the following [23]:
 - The number of elements in the measurement stored in *number_of_elements*.
 - Temperature below which the measured material's electrical resistance drops to 0 (*critical_temp*).
 - The mean, weighted mean, geometric mean, weighted geometric mean, entropy, weighted entropy, range, weighted range, standard deviation, and weighted standard deviation for the following properties of the superconductor—atomic mass, *fi*, radius, density, electron affinity, fusion heat, thermal conductivity, valence. Each of these attributes is labeled using the statistical property first, then underscore, and the property of the superconductor last.

Using these two datasets, the evaluation of the proposed semantic evaluator consists of four main parts: a visualization walkthrough, in which the visual environment and its components are described; qualitative evaluation of the use of large language models in the generation of responses of the prompting; comparative analysis of the approach using two previously designed semantic evaluation models used in correlation analysis, specifically diagnostic sheets [10] and causal heatmaps [11]; and statistical evaluation of the semantic relationship description using Fleiss' κ [24], a model ablation study [25], and ground truth comparisons.

3.1. Visualization Walkthrough

After the loading of the dataset into the implementation of the proposed method, the analyst is presented with an interactive environment containing all the Semantic evaluation cards of the correlation substructures embedded in the data. Figure 7 offers a part of the created environment for the *SuperCond* dataset. As seen, each Semantic evaluation card is presented in its abbreviated form, meaning only the type of substructure (e.g., 4-ptych for n -ptych of $n = 4$), its title using the set of attributes in the substructure, the first sentence of the semantic evaluation, and its visualization miniature are given.

In this environment, the Semantic evaluation cards can be sorted and filtered, which allows analysts to efficiently identify substructures of interest, e.g., those where the relationship is neither causal nor well known in the literature available to the large language model.

Additionally, each of the substructures and their semantic evaluation can be interactively selected for a detailed view of the Semantic evaluation card presented in Figures 8 and 9. As seen, the card is consistent between the two types of substructures with one minor difference in the visualization in the case of hotspot visualization, where the central attribute is colored in a different color to highlight it.

In this case, Figure 8 presents a hotspot substructure for the *AppEn* dataset, where a central attribute $T9$ is visually emphasized and connected to several leaf attributes $T3$, $T4$, and $T7$, illustrating a hub-and-spoke pattern. The accompanying semantic evaluation highlights a strong, consistent relationship in which the central attribute appears to be associated

with the surrounding ones. In contrast, Figure 9 shows an n -ptych of size $n = 5$ for the *AppEn* dataset involving *Press_mm_hg*, *RH_3*, *T1*, *T4*, and *T_out* attributes. The semantic evaluation reflects the complexity of the group-wise relationship and indicates the negative response to the prompting, meaning the absence of a clear, direct causal relationship.



Figure 7. Semantic evaluation card environment for the *SuperCond* dataset.

3.2. Large Language Models in Semantic Evaluation

Figures 8 and 9 show both variants of the semantic evaluation response—a positive one, where the relationship is judged to be well supported by domain knowledge and consistent with known domain knowledge, and a negative one, where the response suggests coincidental correlation that cannot be clearly justified based on available knowledge. This contrast illustrates the range of outcomes produced by the large language model in the semantic evaluation task, demonstrating its ability to both confirm plausible real-world mechanisms and to question or reject unsupported causal interpretations.

In the positive example (Figure 8), the individual prompt responses (#1–#3) consistently identify a clear causal structure, with a central attribute influencing the behavior of the remaining ones. Each response independently highlights a strong correlation between the central vertex and its leaves, interpreting these as indicative of a causal influence consistent with the domain. The final, synthesized analysis uses these responses and combines them into a unified conclusion emphasizing the coherence between the observed correlation-based relationships and the real world.

In contrast, the negative-response example (Figure 9) presents a more complex and densely connected structure, where the large language model responses are more cautious and exploratory. While individual responses acknowledge potential interactions between variables such as temperature, humidity, and pressure, they do not converge on a single,

well-justified cause. Instead, the model emphasizes the possibility of indirect effects, external influences, or coincidental correlations.

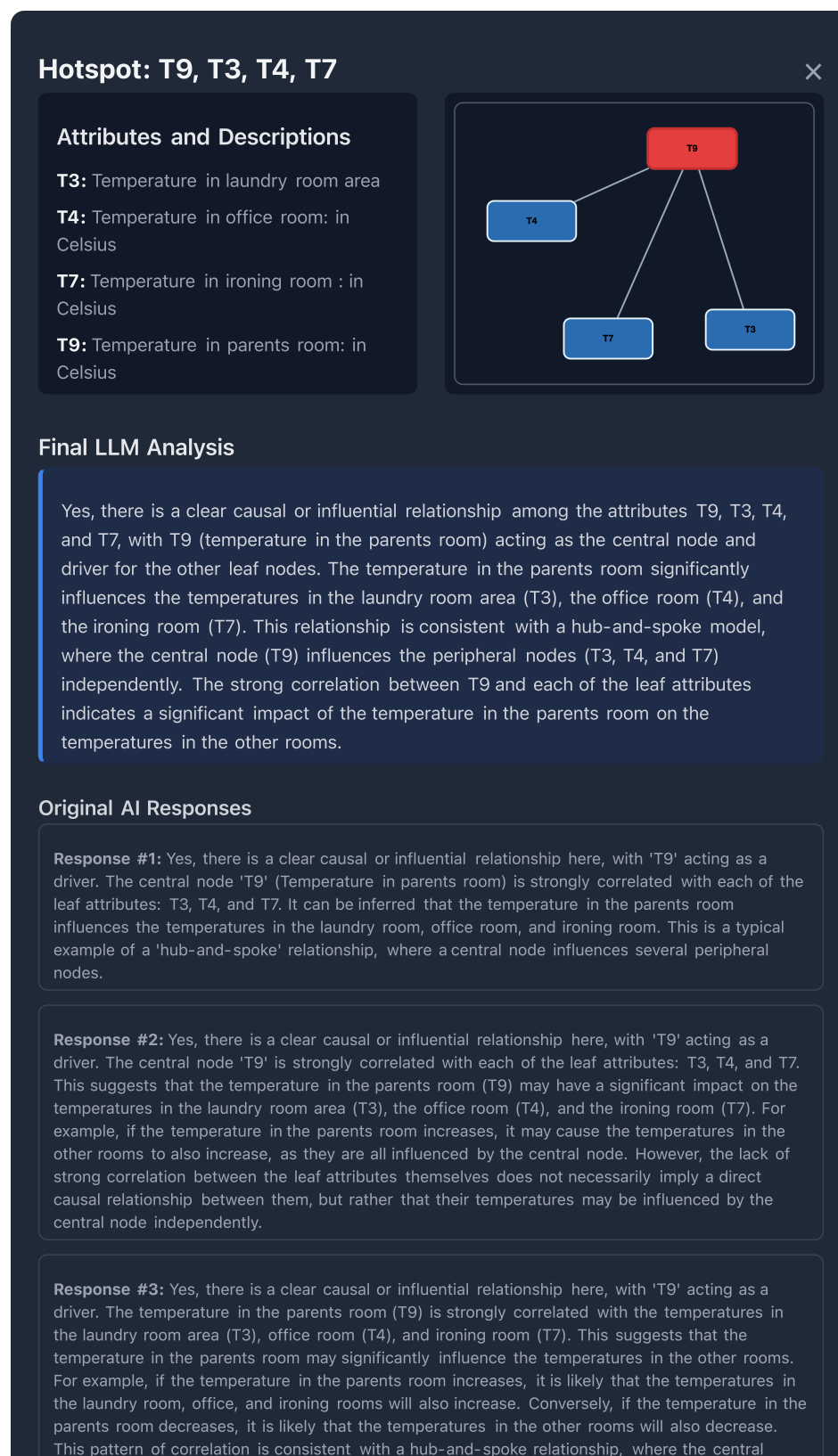


Figure 8. Detailed Semantic evaluation card for a hotspot centered around T9 attribute embedded in AppEn dataset (cropped for the purposes of presentation).



Figure 9. Detailed Semantic evaluation card for a correlation n -ptych of size $n = 5$ embedded in AppEn dataset (cropped for the purposes of presentation).

As noted, each identified substructure is evaluated using three independent prompt responses, which are subsequently synthesized into a final semantic evaluation of the situation. As a result, the total number of large language model interactions scales linearly with the number of substructures, requiring $4 \times \text{number of substructures}$ prompting operations per dataset (three prompts and their synthesis). For datasets with thousands of substructures, this leads to a considerable number of model queries, making optimization via matrix masking strategies an important consideration for practical implementation of the approach.

Table 1 summarizes the number of correlation substructures identified in the correlation graphs of the studied datasets. As shown, the number of embedded substructures is substantial, particularly for the *SuperCond* dataset, where the counts of higher-order n -ptychs reach tens of thousands. Even for the smaller *AppEn* dataset, hundreds of substructures are detected, indicating that exhaustive analysis without prior correlation space pruning would be computationally demanding. Increasing the correlation threshold of σ -masking via the α parameter reduces the number of detected substructures across both datasets significantly, highlighting the need for the proper filtering of the correlation matrices and graphs.

Table 1. Summarization of the number of correlation substructures embedded in the correlation graph of the studied datasets.

Dataset	Number of Attributes	Number of Instances	Number of n -Ptychs			Number of Hotspots
			$n = 3$	$n = 4$	$n = 5$	
$\alpha = 0.1$						
Appliance energy	21	19,735	174	254	252	6
Superconductivity	73	21,263	3554	14,969	48,265	10,248
$\alpha = 0.2$						
Appliance energy	73	21,263	135	150	97	13
Superconductivity	59	21,263	2822	10,741	30,858	5954

In contrast to the substructure-level analysis, Tables 2 and 3 focus on the consistency of the binary judgments of large language models incorporated into the proposed solution and on how the ensemble behaves when individual model responses disagree. Table 2 presents the distribution of unanimous and split decisions, together with Fleiss' κ , offering an overview of the inter-ensemble agreement in the two datasets.

Table 2. Analysis of decision-making of the utilized large language model ensemble.

	AppEn	SuperCond
Unanimous YES	54%	77%
Unanimous NO	16%	2.5%
$2 \times \text{YES} : 1 \times \text{NO}$ split	16%	10.25%
$1 \text{ YES} : 2 \times \text{NO}$ split	14%	10.25%
Fleiss' κ	0.534	0.415

The distribution of responses in Table 2 indicates that the utilized ensemble is skewed toward positive responses, particularly in the *SuperCond* dataset. However, for both of the datasets, the examined relationships are well established and therefore naturally supported by the models. The agreement values measured with the use of Fleiss' κ further confirm this interpretation: an agreement level of 0.534 for the *AppEn* dataset and 0.415

for the *SuperCond* dataset suggest that, although the individual model responses are not aligned, they frequently converge on the same decision. The lower value for *SuperCond* is consistent with a larger number of cases in which the model detects a plausible relationship but remains less certain about its exact semantic support. Overall, the results point to the fact that the ensemble is capable of stable decision-making, while still retaining enough variability to reflect uncertainty in borderline cases.

The comparison with ground truth, supported by the literature focused on the studied datasets [20,22], confirms that the ensemble achieves moderately high decision-making quality on both datasets. Out of 150 randomly selected substructures, the accuracy reaches 86.4% for *AppEn* and 91.1% for *SuperCond*. The main incorrect identifications of relationships semantics in *AppEn* are false positives on noise variables *rv1* and *rv2*, as well as false negatives in physically distant room pairs, where the relationship is indirect and can be influenced by the heating, ventilation, and air conditioning system. In the *SuperCond* dataset, the dominant source of incorrect decision-making is the under-detection of relationships in larger substructures, meaning that more complex relational patterns are harder for the model to resolve consistently.

In addition to the aforementioned quality metrics, an ablation study of individual models in the ensemble was conducted. The results of this ablation, presented in Table 3, provide information on the contribution of individual models to the ensemble outcome. In both studied datasets, the individual large language models (*m1–m3*) are taken out of the ensemble; hence, all pairs of models are evaluated. For the *AppEn* dataset, the highest consistency with the full ensemble is observed for *m3*, which agrees with the ensemble in 98% of decisions, while *m1* shows the lowest agreement at 82%. In terms of accuracy, the combination *m1, m3* performs best among the reduced ensembles, reaching 87.5%, which is even slightly higher than the full ensemble result. This indicates that the third model plays an especially important role in the final decision pattern for the *AppEn* dataset. For *SuperCond*, all ablations remain close to the ensemble baseline, with particularly strong consistency for the combinations *m2, m3* and *m1, m3*. The full ensemble still provides the best performance overall, but the ablation study shows that the ensemble behavior is robust and that no single model alone fully captures the decision pattern.

Table 3. Analysis of decision-making quality for large language model ensemble ablations.

	Model	AppEn	SuperCond
Percentage of decisions consistent with the ensemble	<i>m1</i>	82%	91.1%
	<i>m2</i>	88%	88.6%
	<i>m3</i>	98%	91.8%
	<i>m2, m3</i>	94%	98.8%
	<i>m1, m3</i>	92%	97.5%
	<i>m1, m2</i>	88%	96.8%
Accuracy	<i>m1</i>	68.8%	89.5%
	<i>m2</i>	75%	88.4%
	<i>m3</i>	78%	90.5%
	<i>m2, m3</i>	81.2%	90.5%
	<i>m1, m3</i>	87.5%	91.8%
	<i>m1, m2</i>	81.2%	91.1%
	<i>full ensemble</i>	86.4%	91.1%

However, it should be noted that all three ensemble components correspond to independent responses generated by the same underlying large language model, differing only in the prompt realization rather than in model architecture. Consequently, the ablation analysis does not capture diversity between different models but instead evaluates the effect of aggregating multiple stochastic interpretations of the same semantic task. In this setting, the results indicate that combining three independent evaluations leads to more stable and accurate decision-making than relying on a single response or any reduced two-response configuration, highlighting the benefit of ensemble-style reasoning even within a single-model framework.

3.3. Comparative Analysis of Visualization Approaches to Semantic Evaluation of Correlation Analysis

The final part of the evaluation focuses on the comparative analysis of the proposed approach with three alternative methods for semantic interpretation of correlation structures: the basic manual workflow, the diagnostic sheet visualization presented in [10], and the causal heatmap method introduced in [11]. These approaches differ not only in their level of automation but also in the type of relationships that they analyze and the form of their outputs.

In the manual approach, individual correlation substructures are first identified and visualized separately—the task, which can be done using software tools. However, each of the identified substructures is then semantically evaluated using external knowledge sources such as literature, online materials, or large language models. Naturally, the method is inherently time-demanding and labor-intensive. As the number of substructures increases, this approach quickly becomes impractical and impossible to conduct, especially for high-dimensional datasets.

Diagnostic sheet visualization represents a more structured and automated alternative, and is conceptually closest to the model proposed in this work. Figure 10 visualizes this method applied on the *AppEn* dataset, where the relationship between the attribute pair *T1* and *T8* is selected for large language model semantic analysis. The main difference between the two approaches is that the diagnostic sheet method is limited to pairwise relationships, operating on correlation heatmaps and evaluating only single attribute pairs at a time. This restriction reduces its applicability in more complex analytical scenarios, where relationships among multiple attributes need to be considered simultaneously.

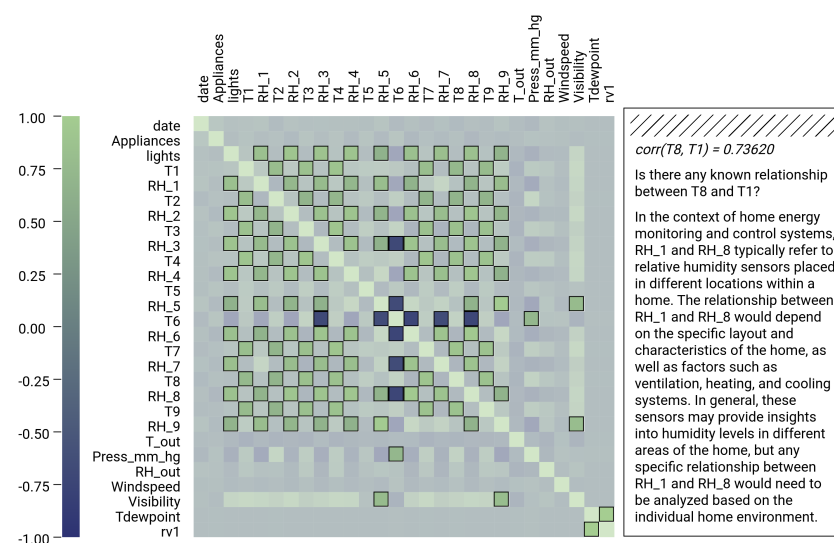


Figure 10. Diagnostic sheet visualization of the *AppEn* dataset (cropped for the purposes of presentation).

The causal heatmap approach introduces a different perspective by incorporating large language model reasoning directly into the correlation matrix through reweighting of its elements. Instead of producing linguistic semantic evaluations, the method modifies correlation coefficients using a predefined familiarity coefficient ω , effectively suppressing coincidental relationships and emphasizing the semantically supported correlations. In Figure 11, such a causal heatmap is presented for the *AppEn* dataset. As can be seen, the visualization focuses more on the strength of the relationship supported by semantic analysis conducted by the large language model than on the linguistic description of relationships in the dataset.

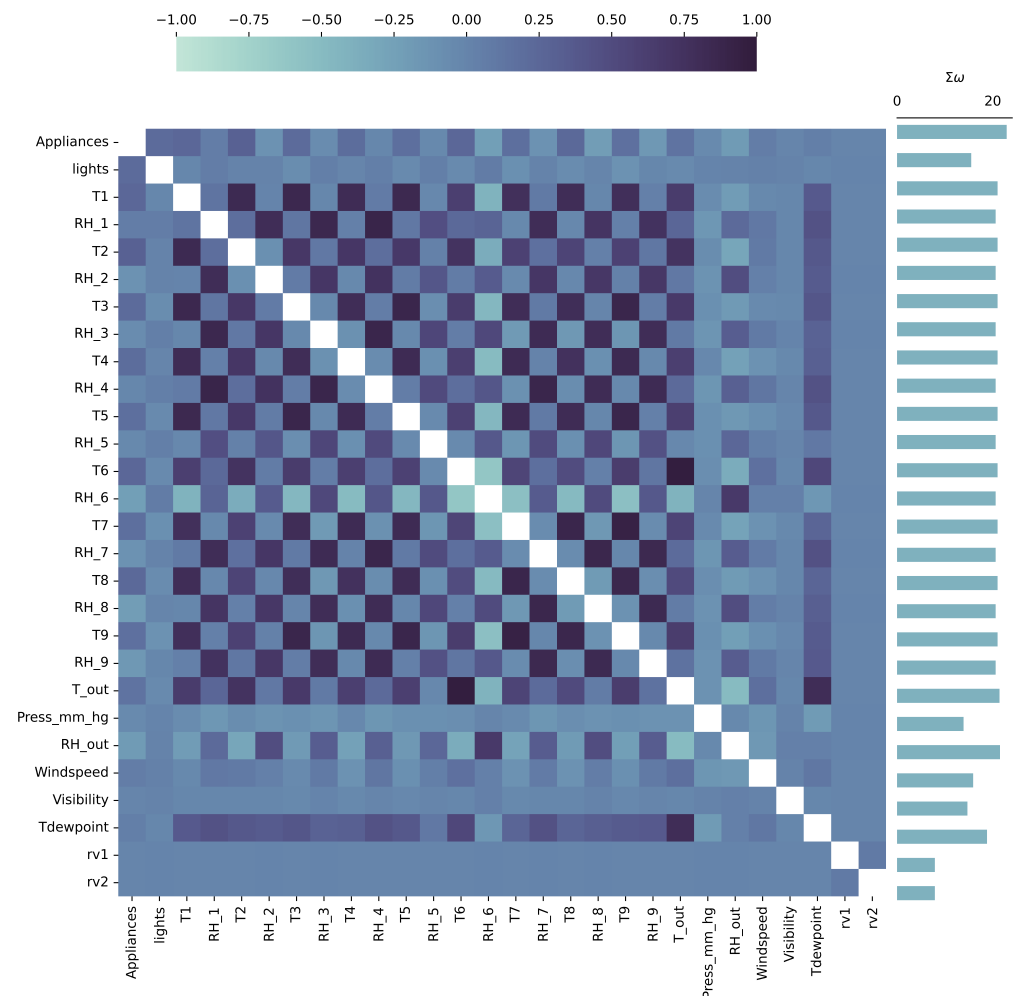


Figure 11. Causal heatmap visualization of the *AppEn* dataset.

As shown in Table 4, the proposed Semantic evaluation cards can be contrasted with the other software-based solutions in several ways—the method focuses on sets of attributes instead of pairs, it supports filtering and sorting, which significantly improves usability when working with large numbers of detected substructures, and features sophisticated three-fold prompting and interactivity with the visual model.

Table 4. Summarization of basic properties of the proposed method in comparison with other semantic evaluation methods used in correlation analysis. (✓—the property is included in the approach, ×—the property is not included in the approach.)

	Diagnostic Sheets	Causal Heatmaps	Semantic Evaluation Cards
Model focus	Correlation of attribute pair	Correlation of attribute pair	Correlation of subset of attributes (substructure)
Output type	Visual, linguistic	Visual, numeric	Linguistic, visual
Output	Correlation map, diagnostic description of attribute pairs	Semantically evaluated correlation map	Correlation substructure and its semantic evaluation
Filtering	×	×	✓
Sorting	×	×	✓
Interactivity	✓	×	✓
LLM model	ChatGPT 4.0	Gemini 2.5	Mistral-7B-Instruct-v0.2
Prompting approach	Direct prompting	Direct prompting with set rigid evaluation structure	Three-fold prompting and synthesis of partial responses
Use case	Semantic analysis of relationship defined by a specific pair of attributes and the value of correlation coefficient measured between them	Modified correlation analysis of a dataset focused on suppressing coincidental (non-causal) and highlighting real-world relationships in data	Semantic analysis of a complex relationship within a specific set of attributes defined by the value of correlation coefficient

3.4. Advantages and Limitations of the Proposed Approach

The proposed Semantic evaluation card environment introduces several notable advantages in comparison to existing approaches for semantic interpretation of correlation structures, namely:

- The method enables the evaluation of entire subsets of attributes rather than restricting the analysis to pairwise relationships. This allows for a more expressive representation of complex, higher-order dependencies within multidimensional datasets, where interactions between multiple variables are often essential for meaningful interpretation.
- Another important advantage is the interactive nature of the proposed environment. The system supports filtering, sorting, and detailed exploration of individual semantic evaluation cards, which was positively evaluated in a preliminary user study conducted with five computer science graduate students. Participants highlighted the clarity of the interface and the usefulness of the interactive features, particularly in terms of navigating large numbers of detected substructures.
- From a quantitative perspective, the large language model-based semantic evaluation achieved an average accuracy of 88.5% across the evaluated datasets, indicating a strong level of agreement with ground-truth interpretations derived from domain literature and dataset documentation.

Despite these advantages, the proposed approach also exhibits two main limitations:

- The primary drawback is its computational complexity, which scales as $O(s(3 + 1)t^2)$, where s denotes the number of embedded substructures and t represents the maximal

number of tokens in the model response. In the conducted experiments, performed locally on an Apple M2 machine (8-core CPU), the evaluation of a single embedded substructure required approximately 45 s. This represents a substantial computational cost when compared to previous approaches. For instance, diagnostic-sheet-based methods exhibit a complexity of $O(n \times t^2)$, where n denotes the number of correlation pairs, while causal heatmap approaches scale as $O(n^2)$ with respect to the number of correlation pairs submitted for analysis. As a result, the proposed method becomes significantly more expensive in scenarios involving large numbers of higher-order substructures.

- Another limitation of the approach is its restricted applicability domain. The method is primarily designed for quantitative attributes and shows reduced reliability when applied to label-encoded categorical variables. Furthermore, it is not yet suitable for time-series data or scenarios involving auto-correlation, where temporal dependencies require specialized modeling techniques beyond the current framework.

4. Conclusions

Semantic evaluation of correlation-based relationships is a fundamental step toward extracting knowledge from multidimensional datasets. Hence, in this work, a novel approach to the process utilizing large language models and interactive visualization called Semantic evaluation cards was introduced. Unlike traditional methods focused on correlation of attribute pairs, the proposed approach enables the analysis of higher-order structures such as correlation n -ptychs and hotspots.

The results of this study show that the presented method effectively distinguishes between relationships that are semantically supported by domain knowledge and coincidental ones, while the interactive card-based environment enhances usability through filtering, sorting, and detailed examination of correlation substructures. However, the approach introduces a notable computational cost, as each substructure requires multiple model interactions, which becomes significant for datasets with large numbers of embedded substructures. When compared to the previously designed approaches, Semantic evaluation cards offer a balanced trade-off between interpretability and analytical depth while reducing the manual effort required in basic workflows, extending beyond the pairwise limitations of diagnostic sheets and providing more interpretable outputs than causal heatmaps.

In the scope of future work in the area, the most significant problem remains the computational costs of the approach, which can be lowered using a combination of parallel computing approaches, the addition of priority to individual substructures, where the substructures will be sent to the large language model in batches based on this priority, and sophisticated correlation structure pruning methods.

Author Contributions: Conceptualization A.D., Methodology A.D. and P.B., Formal analysis A.D., Data analysis A.D. and P.B., Resources A.D., Data curation A.D., Writing—original draft A.D., Writing—review and editing A.D., Visualization A.D. and P.B., Software P.B. All authors have read and agreed to the published version of the manuscript.

Funding: The research presented in this study is supported by the APVV project number APVV-24-0049-Digitization and Intelligent Content Analysis of Historical Documents for Professional Archival Work and Historical Science.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code for the proposed environment for semantic evaluation of correlation structures is available at: <https://github.com/sabaath7766/GPT-Graph-analyzer> (accessed on 10 May 2026). The datasets used in the experimental evaluation of the proposed model are available at: <https://doi.org/10.24432/C5VC8G> and <https://doi.org/10.24432/C53P47>.

Acknowledgments: During the preparation of this manuscript/study, the authors used DeepL 1.89.0, Grammarly 14.1300.0, and ChatGPT 5 for the purposes of grammar and language correction. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dudáš, A. Non-parametric correlation structures and their respective embeddings in predictive analysis. *Discov. Appl. Sci.* **2025**, *7*, 496. [\[CrossRef\]](#)
2. Dudáš, A. Correlation n -ptychs of Multidimensional Datasets. In *Good Practices and New Perspectives in Information Systems and Technologies*; Springer Nature: Cham, Switzerland, 2024.
3. Dudáš, A.; Modrovičová, B. Visualization of Prediction Potential Hotspots in Multidimensional Datasets. *Int. J. Inform. Vis.* **2025**, *9*, 258–264. [\[CrossRef\]](#)
4. Feng, Z.; Liu, J. A Philosophical Analysis of Causality and Correlation—The Debate on Causality between Hume and Bayes. *Comput. Sci. Math. Forum* **2023**, *8*, 13. [\[CrossRef\]](#)
5. Sajjadi Mohammadabadi, S.M.; Kara, B.C.; Eyupoglu, C.; Uzay, C.; Tosun, M.S.; Karakuş, O. A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications. *Electronics* **2025**, *14*, 3580. [\[CrossRef\]](#)
6. Peykani, P.; Ramezanlou, F.; Tanasescu, C.; Ghanidel, S. Large Language Models: A Structured Taxonomy and Review of Challenges, Limitations, Solutions, and Future Directions. *Appl. Sci.* **2025**, *15*, 8103. [\[CrossRef\]](#)
7. Heseltine, M.; von Hohenberg, B.C. Large language models as a substitute for human experts in annotating political text. *Res. Politics* **2024**, *11*, 20531680241236239. [\[CrossRef\]](#)
8. Luo, X.; Rechardt, A.; Sun, G.; Nejad, K.K.; Yáñez, F.; Yilmaz, B.; Lee, K.; Cohen, A.O.; Borghesani, V.; Pashkov, A. et al. Large language models surpass human experts in predicting neuroscience results. *Nat. Hum. Behav.* **2025**, *9*, 305–315. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Musaze, N.; Mezei, J.; Wang, X.L. Exploring the Performance of Large Language Models for Data Analysis Tasks Through the CRISP-DM Framework. In *World Conference on Information Systems and Technologies*; Springer Nature: Cham, Switzerland, 2024.
10. 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, Poprad, Slovakia, 13–15 November 2024.
11. Dudáš, A. Semantic Causality Evaluation of Correlation Analysis Utilizing Large Language Models. *Comput. Mater. Contin.* **2026**, *87*, 76507. [\[CrossRef\]](#)
12. Zhang, X.; Xu, C.L.; Lu, Z.X.; Wang, H.N.; Song, D. ReBaCCA-ss: Relevance-Balanced Continuum Correlation Analysis With Smoothing and Surrogating for Quantifying Similarity Between Population Spiking Activities. *Neural Comput.* **2026**, *38*, 651–680. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Cai, J.H.; Zeng, W. A hierarchical Bayesian latent class mixture model with censorship for detection of linear changes and correlation analysis across populations in antimicrobial resistance. In *Statistical Theory and Related Fields*; Taylor & Francis: Abingdon, UK, 2026; pp. 1–12.
14. Henry, C.J.; Awais, S.A. A Descriptive Tolerance Nearness Measure for Performing Graph Comparison. *Fund. Informaticae* **2018**, *163*, 305–324. [\[CrossRef\]](#)
15. Sun, R.J.; Chen, C.; Wang, X.Y.; Wu, Y.P.; Zhang, W.J.; Zhang, Y.; Lin, X.M. Efficient maximum signed biclique and biplex identification in signed bipartite graphs. *VLDB J.* **2024**, *34*, 60. [\[CrossRef\]](#)
16. Feng, W.B.; Hélie, S.; Panchal, J.H. Enhancing design concept diversity: Multi-persona prompting strategies for large language models. *Des. Sci.* **2025**, *11*, e52. [\[CrossRef\]](#)
17. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825. [\[CrossRef\]](#)
18. Kumar, P. Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artif. Intell. Rev.* **2024**, *57*, 260. [\[CrossRef\]](#)
19. Wang, L.P.; Chen, S.; Jiang, L.N.; Pan, S.; Cai, R.Z.; Yang, S.; Yang, F. Parameter-efficient fine-tuning in large language models: A survey of methodologies. *Artif. Intell. Rev.* **2025**, *58*, 227. [\[CrossRef\]](#)

20. Candanedo, L.M.; Feldheim, V.; Deramaix, D. Data driven prediction models of energy use of appliances in a low-energy house. *Energy Build.* **2017**, *140*, 81–97. [CrossRef]
21. Candanedo, L. Appliances Energy Prediction [Dataset]. UCI Machine Learning Repository. 2017. Available online: <https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction> (accessed on 10 May 2026).
22. Hamidieh, K. A data-driven statistical model for predicting the critical temperature of a superconductor. *Comput. Mater. Sci.* **2018**, *154*, 346–354. [CrossRef]
23. Hamidieh, K. Superconductivity Data [Dataset]. UCI Machine Learning Repository. 2018. Available online: <https://archive.ics.uci.edu/dataset/464/superconductivity+data> (accessed on 10 May 2026).
24. Liu, C.; Chen, Y.; Xao, X. Tracing Cognitive Dynamics in Group Ideation: An LLM-Based Framework for Automating Linkography. *Int. J. Hum.-Comput. Interact.* **2026**, *146*, 1–19. [CrossRef]
25. Debelie, A.; Bagui, S.S.; Bagui, S.C.; Mink, D. A Systematic Ablation Study of GAN-Based Minority Augmentation for Intrusion Detection on UWF-ZeekData22. *Electronics* **2026**, *15*, 1291. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.