

Explaining Complex ML Models to Domain Experts Using LLM & Visualization: An Exploration in the French Breadmaking Industry

Briggs Twitchell
Northeastern University
Portland, Maine, USA
twitchell.b@northeastern.edu

Anastasia Bezerianos
LISN, Université Paris-Saclay, CNRS, INRIA
Orsay, France
anastasia.bezerianos@universite-paris-saclay.fr

George Katsirelos
MIA Paris-Saclay
INRAE
Paris, France
georgios.katsirelos@inrae.fr

Nadia Boukhelifa
UMR 518 MIA-PS, Université Paris-Saclay
INRAE
Palaiseau, France
UAR 3611 CNRS
ISC-PIF
Paris, France
nadia.boukhelifa@inrae.fr

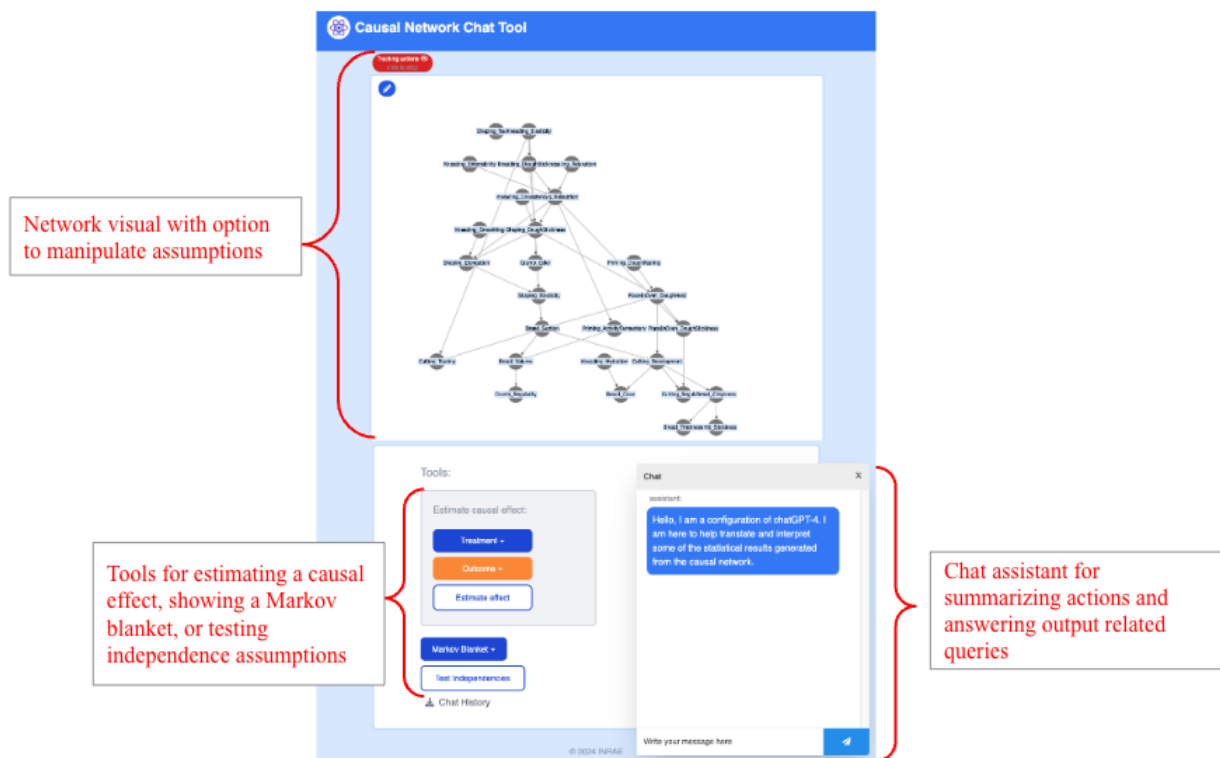


Figure 1: The full visualization application with its network visualization, suite of tools, and chat assistant.

Abstract

Modeling a complex system from data can aid understanding and decision-making. Bayesian networks are one such method that,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI EA '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1395-8/25/04
<https://doi.org/10.1145/3706599.3706685>

when accurately constructed, can support inference and help understand the underlying system that generated the data. However, the outputs of these models are not always intuitive, especially for users that lack a statistical background. In this work, we examine how the recent advancements in modern Large Language Models (LLMs) may be applied to help explain machine learning (ML) models. Following a user-centered design methodology, we collaborated with a team of ML modelers and a domain expert in the French breadmaking industry to develop a causal inference application with an integrated chat assistant. From qualitative feedback sessions with modelers and the domain expert, we note some unique advantages but also a host of challenges in using current LLMs for model explainability.

CCS Concepts

• **Human-centered computing** → **Empirical studies in visualization**; **Visual analytics**.

Keywords

visualization, working with domain experts, Bayesian networks, modeling, bread-making.

ACM Reference Format:

Briggs Twitchell, George Katsirelos, Anastasia Bezerianos, and Nadia Boukheifa. 2025. Explaining Complex ML Models to Domain Experts Using LLM & Visualization: An Exploration in the French Breadmaking Industry. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*, April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3706599.3706685>

1 Introduction

In several domains, including agronomy, food science and manufacturing, domain experts often use models of complex processes, in order to understand underlying relationships in their data, reach insights and make decisions about how to optimize these processes. For example, projects like EVAGRAIN [1] attempt to model data collected from thousands of breadmaking tests, in order to create decision support tools. Domain experts can later use these tools to investigate relationships between wheat production cost, wheat type, bread quality and nutritional value, balancing concerns around nutrition and sustainable wheat production [25].

One way to model such processes are Bayesian Networks (BNs). BNs are normalized, decomposed graphical representations of a *probability distribution* over a set of variables [27]. In the case of the breadmaking process for example, these variables can be sensory measurements from different phases of the breadmaking process such as dough elasticity or hydration (i.e. liquid content).

BNs are typically represented as Directed Acyclic Graphs (DAGs), where vertices are a set of nodes, each representing a variable (often called "random variables" as they follow a joint probability distribution). Links between nodes form a set of directed edges, representing conditional dependencies between the variables (for example, the dependency between dough hydration and dough elasticity). Each node in the graph has an associated conditional probability distribution, which quantifies the effect of the parent nodes on this node. For example, if the parent node "dough hydration" has a high value, then the child node "dough slackening" is

very likely to have a high value. The joint probability distribution over all variables in the network is given by the product of the conditional probabilities of each node given its parents.

BNs are a form of Machine Learning (ML) and because they can be presented graphically, they can be considered as explainable in the domain of eXplainable Artificial Intelligence (XAI) [9]. Despite this relative transparency, there still exist challenges in communicating information produced by BNs to non-technical end-users, especially when networks contain a large number of variables [21]. Additionally, BNs, encode probabilistic associations among variables in the form of conditional (in)dependencies, which may lead to confusion, as humans tend to reason in causal terms [16, 39].

As is discussed in the following section, there are many different methods which attempt to explain model outputs. In this work, we explore how recent advancements in Large Language Models (LLMs) may offer a useful means for non-technical users to develop and interpret ML models. Following a user-centered design methodology, and alongside a team of modelers in the breadmaking industry, we developed an application that integrates ChatGPT-4 to assist in the explanation of various model operations. We discuss lessons learned about integrating the chat service and how it was perceived by domain experts. Our findings can help guide the design of explainable user interfaces for machine learning models, and more specifically Bayesian networks.

2 Related Work

In the past several years, machine learning (ML) models have seen notable improvements in performance, largely attributable to models of greater size and complexity. In response to this growing complexity, the field of XAI emerged as a collection of methods for understanding these models. Certain domains, such as medicine and law, demand reasons for decisions, and thus the application of ML models requires a degree of transparency and trust [2]. Popular XAI methods include Local interpretable Model-agnostic Explanations (LIME) [32], SHapley Additive exPlanations (SHAP) [23], partial dependence plots [15] and pixel saliency maps [34, 42], each of which seeks to quantify input feature importance in some fashion. For Bayesian Networks (BNs) specifically, explanations most often focus on either explaining the evidence or on explaining the model's structure [9, 21].

Approaches focusing on explaining evidence try to communicate the state of the variables of the BN that are most consistent with this evidence. Given some evidence (for example, that bread dough elasticity is low) there are approaches that search for the Most Probable Explanation (MPE) and Maximum Posterior Probability (MAP), in other words they search for the most likely configuration of all other unobserved variables [18] (for example, if the observed dough elasticity is low, what are the most likely values for all other variables such as dough hydration, shape, or volume). Extensions of this approach, such as Most Relevant Explanation (MRE), narrow in on a subset of explanatory variables related to the observed evidence that communicate the most information [41]. Likewise, Flores [12] describes an explanation tree construction algorithm to maximize information while giving preference to conciseness. The BN model structure is often explained visually. GUI-based applications have long been used to visualize network structure as a graph, combined

with probability tables (e.g., [3, 4, 20]). Some visual approaches (e.g., [37]) allow for visual *evidence propagation*, in other words, users can set the values of some nodes of the network (e.g., dough elasticity to low) and see visually the values and probabilities change for the rest of the network.

Another user-oriented approach has been the development of natural language explanations. Vlek et al. [40] propose a scenario-based method to generate narrative-based reasoning, and Keppens [17] uses support graphs to produce explanatory text in natural language, given an evidence set. Recent work by Kiciman et al. [19] examines the capacity of ChatGPT-3 and 4 for causal reasoning and discovery. However, to our knowledge, the use of modern LLMs to explain inference to non-technical users has yet to be explored and thus helped guide the focus of this project.

Additive feature value explanations such as SHAP exhibit alignment with human intuition [23]. However, they apply only to understanding models and are thus incomplete for the task of understanding the underlying causal relationships among variables [10]. The explanation methods specific to BNs - MPE, MRE, and explanation trees - focus primarily on the contents of explanations, rather than the format. The works by [17, 20, 36, 40] describe different means by which this content may be presented to end-users, but do not assess effectiveness from a user perspective.

End-user profile is an important factor influencing the effectiveness of explanation media, whether numeric, visual, linguistic, or a combination thereof. Szymanski et al. find that non-expert users benefit from textual explanations, while, experts, (i.e. those more familiar with the data and domain) obtain little to no benefit [35]. Bertrand et al. note that, generally speaking, expert users approach explanations more critically, while lay users may over rely on natural language explanations, posing a potential mismatch between user trust and actual understanding [5].

Szymanski et al. do find that hybrid explanations - both graphical and textual - significantly improve non-expert user understanding and trust [35], though Bertrand et al. come to a different conclusion - that graphical representations alone are preferable [5]. This application implements a hybrid approach wherein the chat assistant acts as a supplement to the DAG visual. The feedback we obtained focused on the utility of this supplemental component.

3 Application Domain and Explainability Requirements

For this project we collaborated with three modelers/ML experts having on average over five years of modelling experience in bread-making, and one agricultural researcher (domain expert), with over 10 years of experience in the bread-making process (more precisely acting as a 'proxy' for domain experts in the French breadmaking industry). The dataset used in the application contained 27 sensorial measurements, each of which could take on up to seven possible values: very insufficient, insufficient, slightly insufficient, normal, slightly excessive, excessive, and very excessive. To simplify the network, these measurements were condensed to a maximum of three possible values: insufficient, normal, excess. There were 287 total observations. Domain experts identified six sequential phases in the breadmaking process : *kneading*, *first rising*, *dividing*, *second rising*, *baking*, and *bread analysis* [24].

The network's initial structure was learned via the Greedy Hill Climbing algorithm guided by Bayes Information Criterion, respecting temporal restrictions imposed by the aforementioned stages - for example, measurements obtained when the bread is placed in the oven could not causally influence measurements in the kneading stage. Thus, the network reflected a combination of algorithmically learned patterns and domain expert knowledge. For greater detail on the dataset or network's construction, see [24].

We conducted an hour long general brainstorming session partially in-person and partially via Zoom to determine user requirements with regards to explainability of the BN model. There were four participants with a range of different backgrounds. One participant was an expert in Bayesian networks, one participant was a researcher studying modeling for breadmaking, and two participants were agricultural researchers from the same institute. This discussion yielded two primary takeaways:

- The first was an emphasis on model interpretability: users would likely give little consideration to a model for which they did not understand and could not reconcile to their own knowledge of the breadmaking process.
- The second was the development of a model that could help identify the cause of non-desirable instances in the dataset - for instance, dough with an inability to maintain its shape during the kneading stage.

This former concern regarding user trust in the model has been documented, mostly in the context of the medical domain [7, 31]. It oriented our focus toward improving model transparency for end-users. The latter concern regarding causal explanations directed us to explore applications of causal inference methods.

4 Approach and Visualization Platform

The application we developed constructs a graphical display of a causal model for the breadmaking process and permits the following operations:

- **Graph manipulation:** the user may click on nodes to create or delete edges, representing the existence or non-existence of directional causal assumptions.
- **Do-calculus:** the user may simulate the effect of intervening on a target variable by identifying an adjustment set and estimating a causal effect [28].
- **Markov blanket visualization:** the user may select a target variable and display its Markov Blanket. For a given node X , the Markov blanket M_X is defined as the set of nodes with direct edges pointing to X , direct edges originating from X , or direct edges pointing to children of X . Given a node's Markov blanket, no additional information will change this node's expected value [18].
- **Independence testing:** the user may evaluate the independence assumptions present in the graph via a chi2 test of all minimal conditional independencies. This test returns a list of the p-values from each test.

The results from one or several of the above operations are used as context for the chat assistant, which was implemented

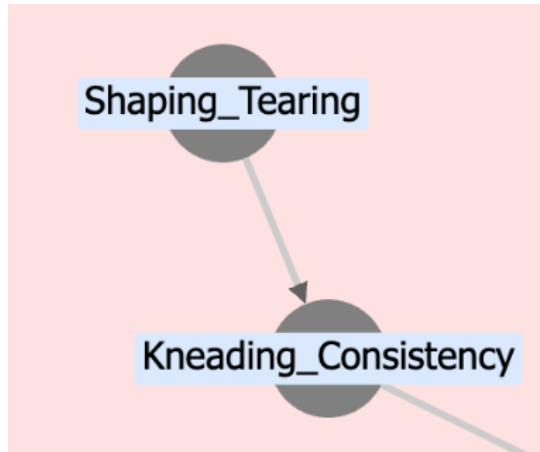


Figure 2: The user may force a causal assumption or independence between two variables – this instance represents a causal relationship flowing from *Shaping_Tearing* to *Kneading_Consistency*

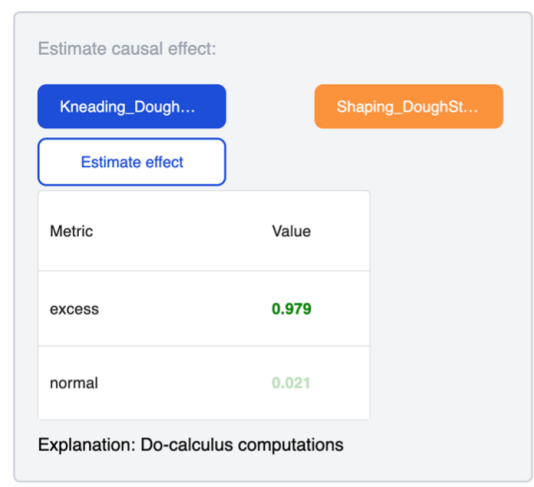


Figure 3: A causal estimate for a treatment variable *Kneading_DoughStickiness* and outcome variable *Shaping_DoughStickiness*

with the help of the OpenAI API and instantiated as an instance of ChatGPT-4 [26]. The application uses Cytoscape [13] and React [30] for graph visualization and frontend rendering, and it uses the pyAgrum library [11] for all backend causal model operations.

These components work together to help address the two requirements identified in the brainstorming session. The do-calculus tool, graph manipulation tool, Markov blanket visualization, and independence assumption testing permit users to assess how the variables causally relate to one another, which may help them identify key variables contributing to non-desirable outputs. The chat assistant summarizes these outputs in natural language and respond

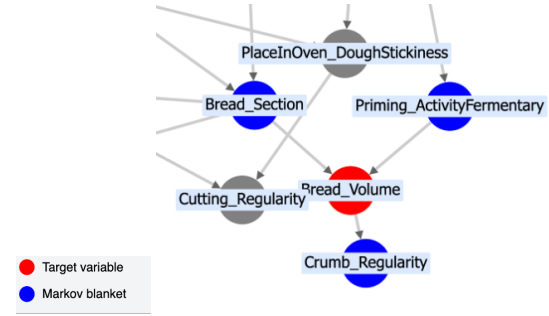


Figure 4: The Markov blanket of a target variable *Bread_Volume* represented by nodes with blue color.

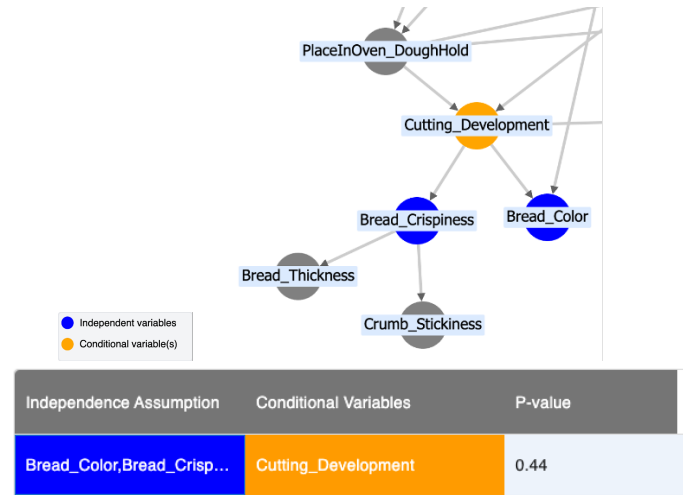


Figure 5: Independence assumption test of the variables *Bread_Color* and *Bread_Crispiness* when conditioned on *Cutting_Development*, with a resulting p-value of 0.44.

to specific queries by the user for the purpose of making the model more interpretable¹.

5 Qualitative Validation and Results

We obtained qualitative feedback from the three modelers and the domain expert in three sessions, each roughly an hour long, conducted via Zoom. The first two sessions consisted of presenting the application to the participants, each working individually, and where live feedback was obtained regarding application use cases and potential improvements. Feedback from these sessions was used to improve the visualization platform and suite of tools and operations implemented. The final session consisted of two participants demoing the application themselves (one domain expert working with one modeler), and another modeller interacting with the tool separately. Participants then gave joint feedback on the

¹In this paper, we define "interpretability" as the transparency of a model (for example, the causal estimates that conform to the underlying assumptions and dataset), and we define "explainability" as the effectiveness by which a model is made transparent via some post-hoc method [2]

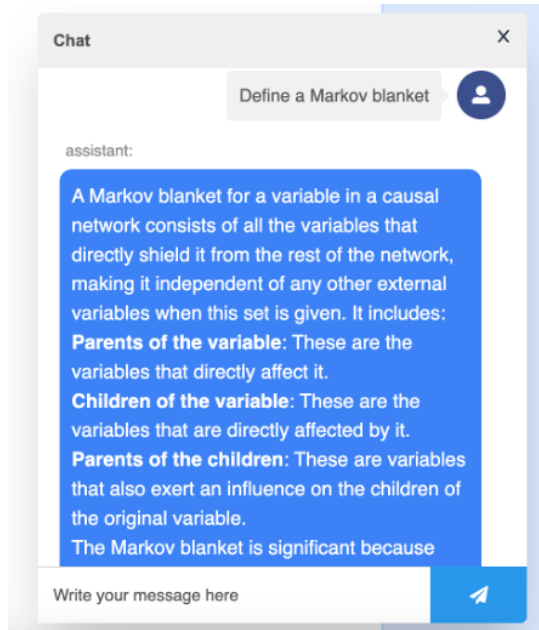


Figure 6: An example response from the chat assistant, which summarizes outputs from user interactions and answers user queries, as provided above.

platform and the different functionalities, before responding to an online questionnaire examining the utility of the visualization and causal inference platform.

More specifically, in this questionnaire we asked participants which functionalities in the platform they found most helpful for understanding the network; whether the chat feature helps their understanding of the causal inference operations performed on the network; whether the chatbot’s output appears consistent with the network, with the participants’ familiarity with the dataset and use case; and finally what features/operations/visualizations they would like to add to the application.

- **Preference for simplicity and conciseness:** The modelers noted that the output from the chat assistant were overly detailed and long-winded. They tended to disregard its explanations when they were verbose and technical, as it required too much effort. One participant commented: “The answers from the chat are too long for me. I kind of can’t be bothered to read so many words. And honestly, with the knowledge on the meaning of each computation, there would be no need for the complete explanation”. This finding is in line with Rosenfield’s work [33], which identifies conciseness as a key element of good explanations.
- **Relevance of the visualization and GUI:** Much of the feedback for the application concerned the presentation of the selected operations and network – for instance, the menu selectors and colors used for the Markov blanket visualization. This finding exhibits difficulty in assessing the utility of the chat assistant independent of the other features of the application and also suggests that, without a well-designed

user interface, even good textual explanations will be underutilized by the user. In particular, one participant wanted to see more visual representations related to the chatbot output: “It is hard to make sense to certain results, maybe a visualization of the values (+ number of value of the dataset) for a given variable can help to understand certain”. This potentially calls for integrating chatbot output into the main visualization (graph view).

- **Chat output and variability:** We noted mixed results regarding the usefulness of the chat assistant. Two participants noted that it displayed some utility in documenting and summarizing user interaction history, for example, one participant noted: “Chat is especially useful to keep track of the tests and summarize the result, however it generates way too much verbose”. However, this varied depending on the operations to which it was responding. It displayed an ability to offer general background information for an operation’s output. However, for elements of the prompt that were more specific to the breadmaking domain, the responses were not useful. For example, “Generally yes but there are inconsistencies in the text, especially when the interpretation goes too deep about the nature of the variable (e.g. controlled versus uncontrolled factors)”. Below is an example that illustrates this:

Prompt: The assumption that the variables *Bread_Color* and *Bread_Thickness* are independent when conditioned on [*Bread_Crispiness*] has a *p*-value of 0.

Response: The test you’ve conducted provides strong evidence against the assumption that *Bread_Color* and *Bread_Thickness* are conditionally independent given *Bread_Crispiness*. A *p*-value of 0 indicates there is no statistical support for independence, suggesting that even after accounting for *Bread_Crispiness*, there remains a linkage or interaction between *Bread_Color* and *Bread_Thickness*. In practical terms for breadmaking, this suggests:

- The color and the thickness of the bread are related, and this relationship is not entirely explained by the crispiness of the bread.
- Factors contributing to either the thickness or color of the bread might still affect the other, irrespective of the bread’s crispiness.

Adjustments in your breadmaking process that aim to alter the color might unintentionally influence thickness, and vice versa, not merely through their impact on crispiness. This insight can be important when attempting to independently control these qualities during bread production.

Though the chat assistant’s response accurately interprets the results of the test, which from an educational perspective may be helpful to users unfamiliar with statistical tests, it is long and comments extensively on the test’s implications. The users regarded these responses as overly verbose.

6 Discussion, Lessons Learned and Future Work

Though the application managed to provide insightful analysis in certain instances, there are a host of factors that will require

improvement before it proves useful to end-users. We report next on the three most important lessons learned:

Challenge of Domain-Specific Queries: The chat assistant was capable of answering general queries about the application’s operations, but prompts that were more specific either to the domain of breadmaking or the state of the network yielded sub-optimal responses. We also noted that the quality of chat assistant responses was highly sensitive to the textual representation of the graph, operation performed, and naming of variables (for instance, using English or French variable names).

Power and Drawbacks of Dialogue: A benefit of the chat assistant, as of yet unexplored in previous work on this subject, is the ability for users to ask follow-up questions to the chat responses, adding an element of interaction and flexibility. This extends natural language explanations to a suite of diverse operations, as implemented here. However, with this flexibility also comes a lack of predictability, and thus this approach is not directly comparable to the prior natural language generation methods mentioned, which constrain their focus to an explanation of observed evidence. We also note that the operations selected for this application are by no means exhaustive and instead serve as a starting point to assess the utility of a chat assistant in this context.

Advanced Approaches to Enhance LLM Integration: In this work, we considered simple prompts to interact with the chatbot, often leading to overly detailed and long-winded responses. Future work could focus on leveraging diverse LLMs, such as Gemini [8] and LLaMA [38], alongside ChatGPT to enable performance comparisons including quantitative validation. Incorporating advanced techniques like domain-specific prompt engineering and retrieval-augmented generation [22] could improve the LLM’s response quality. Additionally, a systematic approach to comparing successful and unsuccessful LLM interactions would provide further insights on how to more effectively integrate LLMs for ML model explainability.

Through our discussions with our experts (modelers and domain experts), we also identify avenues for future research and considerations for the field:

- **Autonomous Operations by Chat Assistants:** Enable the chat assistant to autonomously perform causal model operations and return results to the user. For instance, the user could input via text to the chat assistant a desired modification to the causal structure of the network. The assistant could respond by calling a series of functions to modify the network, re-render it to the user, and finally summarize some implications of this change. Under this approach, user interactions with the application would be the mirror image to the approach taken in this paper – that is, instead of interacting directly with the DAG or causal inference tools to create text summaries of the results, users could input text and view these effects on the DAG and tools.
- **Local Explanations:** Develop capabilities for the chat assistant to provide explanations for specific observations or data points. This feature would allow for more targeted and

practical insights, particularly useful for addressing specific anomalies or points of interest in the data. For example, Budhathoki et al [6] document an approach for root cause identification that could be applied in the breadmaking process to determine the variables producing undesirable observations.

- **Graph Comparison:** The users expressed a desire to obtain scores for DAG accuracy as well as compare altered graphs to previous versions. This could permit users to track the evolution of the graph over time. Methods from comparative visualization [14] can be used to facilitate this comparison, taking into account various sources of information that contributed to the construction of the specific version of that model (e.g., data, machine learning algorithm, user expertise and the LLM).
- **Filtering:** Domain experts are often familiar with just a subset of variables in the network, as for instance they may have expertise in one part of the breadmaking process. Thus allowing them a means to filter chat responses to specific parts of the network could narrow their analysis on the area of their domain knowledge. In our application, we used the log of user interactions to constrain the context of the LLM, but we could also consider the same log data (i.e. analytical provenance [29]) to also filter to relevant data, make suggestions of related variables and / or atypical outcomes, and to ask and consider different user goals such as understanding the model versus debugging it.
- **Impact on Decision Making:** Our work suggests that a chat assistant can prove a useful tool for explaining models, but it is unclear if these explanations may bias user understanding and in the long-run decision making. Our preliminary results indicate that domain-specific responses from the chat were sub-optimal, this could impact experts in different ways. It may lead them to uninformed decisions if they trust the model and chat. Or it could reduce their trust not only in the chatbot output, but potentially their trust on the model if they do not understand the distinction between the two. Studying this impact remains future work.

7 Conclusions

In summary, the integration of modern LLMs into causal inference applications offers a new perspective for improving model interpretability. However, our study highlights challenges, particularly with domain-specific queries, the strengths and limitations of dialogue with the chatbot, and the need for more advanced approaches to enhance LLM Integration. Additionally, it remains unclear how to balance the level of detail—in both chatbot output and the visualization/GUI—while ensuring the explanations are understandable for the end user and remain relevant to the domain.

Acknowledgments

This research was supported by the DATAIA convergence institute as part of the "Programme d’Investissement d’Avenir" (ANR- 17-CONV-0003), and the ANR Evagrain project (ANR-20-CE21-0008). We thank Pierre-Henri Willemin for his assistance with pyAgrum, Mélanie Münch, Cédric Baudrit, and Kamal Kansou for the helpful discussions, and the reviewers for their insightful comments.

References

- [1] 2020. ANR EVAGRAIN project. <https://anr.fr/Project-ANR-20-CE21-0008>. Accessed: 2024-09-20.
- [2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Benetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [3] Bayes Server. 2024. Bayes Server - Advanced Analytics for Bayesian Networks. <https://www.bayesserver.com/> Accessed: 2024-06-05.
- [4] BayesFusion. 2024. BayesBox - Interactive Bayesian Network Tool. <https://demo.bayesfusion.com/bayesbox.html> Accessed: 2024-06-05.
- [5] Astrid Bertrand, James R. Eagan, and Winston Maxwell. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 943–958. <https://doi.org/10.1145/3593013.3594053>
- [6] Kailash Budhathoki, Dominik Janzing, Patrick Bloebaum, and Hoiyi Ng. 2021. Why did the distribution change?. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1666–1674.
- [7] Raphaela Butz, Renée Schulz, Arjen Hommersom, and Marko van Eekelen. 2022. Investigating the understandability of XAI methods for enhanced user experience: When Bayesian network users became detectives. *Artificial Intelligence in Medicine* 134 (Dec. 2022), 102438. <https://doi.org/10.1016/j.artmed.2022.102438>
- [8] Google DeepMind. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>.
- [9] Iena Petronella Derks and Alta de Waal. 2020. A Taxonomy of Explainable Bayesian Networks. In *Artificial Intelligence Research (Communications in Computer and Information Science)*, Aurora Gerber (Ed.). Springer International Publishing, Cham, 220–235. https://doi.org/10.1007/978-3-030-66151-9_14
- [10] SHAP Documentation. 2018. Be careful when interpreting predictive models in search of causal insights. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html. Accessed: 2024-06-23.
- [11] G. Ducamp, C. Gonzales, and P.-H. Wuillemin. 2020. aGrum/pyAgum: a Toolbox to Build Models and Algorithms for Probabilistic Graphical Models in Python. In *10th International Conference on Probabilistic Graphical Models (Proceedings of Machine Learning Research, Vol. 138)*. Skørping, Denmark, 609–612.
- [12] M. Julia Flores, José A. Gámez, and Serafin Moral. 2005. Abductive Inference in Bayesian Networks: Finding a Partition of the Explanation Space. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Lluís Godó (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 63–75.
- [13] Max Franz, Christian T. Lopes, Gerardo Huck, Yue Dong, Onur Sumer, and Gary D. Bader. 2016. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32, 2 (Jan. 2016), 309–311. <https://doi.org/10.1093/bioinformatics/btv557>
- [14] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. 2011. Visual comparison for information visualization. *Information Visualization* 10, 4 (2011), 289–309.
- [15] Brandon M. Greenwell, Bradley C. Boehmke, and Andrew J. McCarthy. 2018. A Simple and Effective Model-Based Variable Importance Measure. <https://doi.org/10.48550/arXiv.1805.04755> arXiv:1805.04755 [cs, stat].
- [16] Conor Hennessy, Alberto Bugarin, and Ehud Reiter. 2020. Explaining Bayesian Networks in Natural Language: State of the Art and Challenges. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, Jose M. Alonso and Alejandro Catala (Eds.). Association for Computational Linguistics, Dublin, Ireland, 28–33. <https://aclanthology.org/2020.nl4xai-1.7>
- [17] Jeroen Keppens. 2019. Explainable Bayesian Network Query Results via Natural Language Generation Systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. ACM, Montreal QC Canada, 42–51. <https://doi.org/10.1145/3322640.3326716>
- [18] D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. <https://books.google.co.in/books?id=7dzpHCHzNQ4C>
- [19] Emre Kicman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. <https://doi.org/10.48550/arXiv.2305.00050> arXiv:2305.00050 [cs, stat].
- [20] Carmen Lacave, Roberto Atienza, and Francisco J. Diez. 2000. Graphical Explanation in Bayesian Networks. In *Medical Data Analysis*, Rüdiger W. Brause and Ernst Hanisch (Eds.). Springer, Berlin, Heidelberg, 122–129. https://doi.org/10.1007/3-540-39949-6_16
- [21] Carmen Lacave and Francisco J. Diez. 2002. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review* 17, 2 (June 2002), 107–127. <https://doi.org/10.1017/S026988890200019X> Publisher: Cambridge University Press.
- [22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [23] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. <https://doi.org/10.48550/arXiv.1705.07874> arXiv:1705.07874 [cs, stat].
- [24] Mélanie Munch, Cédric Baudrit, Hubert Chiron, Benoît Méléard, Luc Saulnier, and Kamal Kansou. 2024. Diagnosis based on sensory data: Application to wheat grading quality. *Innovative Food Science & Emerging Technologies* 96 (2024), 103771.
- [25] Melanie Munch, Cedric Baudrit, Kamal Kansou, and Christophe Fernandez. 2023. Conception d'un Outil de Diagnostic: Application à l'Essai de Panification en Industrie Boulangère. In *JFRB 2023: 11èmes Journées Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes*. <https://hal.science/hal-04190423/>
- [26] OpenAI. 2024. OpenAI Assistants. <https://platform.openai.com/assistants> Accessed: 2024-06-05.
- [27] Judea Pearl. 1995. From Bayesian Networks to Causal Networks. In *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, Giulianella Coletti, Didier Dubois, and Romano Scozzafava (Eds.). Springer US, Boston, MA, 157–182. https://doi.org/10.1007/978-1-4899-1424-8_9
- [28] Judea Pearl. 2009. *Causality* (2nd ed.). Cambridge University Press.
- [29] Eric D Ragan, Alex Endert, Jibonanda Sanyal, and Jian Chen. 2015. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 31–40.
- [30] React. 2024. React - A JavaScript library for building user interfaces. <https://reactjs.org/> Accessed: 2024-06-05.
- [31] Ehud Reiter. 2019. Natural Language Generation Challenges for Explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, Jose M. Alonso and Alejandro Catala (Eds.). Association for Computational Linguistics, 3–7. <https://doi.org/10.18653/v1/W19-8402>
- [32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. <https://doi.org/10.48550/arXiv.1602.04938> arXiv:1602.04938 [cs, stat].
- [33] Avi Rosenfeld. 2021. Better Metrics for Evaluating Explainable Artificial Intelligence. In *Adaptive Agents and Multi-Agent Systems*.
- [34] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. <https://doi.org/10.48550/arXiv.1706.03825> arXiv:1706.03825 [cs, stat].
- [35] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 109–119. <https://doi.org/10.1145/3397481.3450662>
- [36] Johannes Textor, Benito van der Zander, Mark S Githorpe, Maciej Liśkiewicz, and George TH Ellison. 2016. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology* 45, 6 (Dec. 2016), 1887–1894. <https://doi.org/10.1093/ije/dyw341>
- [37] Pierpaolo Tommasi, Stéphane Deparis, and A. Pascale. 2019. HWPProfile UI: facilitating the exploration of a patient centered risk model. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. 1–2. <https://doi.org/10.1109/ICHI.2019.8904573>
- [38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [39] Amos Tversky and Daniel Kahneman. 1977. Causal Thinking in Judgment under Uncertainty. In *Basic Problems in Methodology and Linguistics: Part Three of the Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science, London, Ontario, Canada-1975*, Robert E. Butts and Jaakko Hintikka (Eds.). Springer Netherlands, Dordrecht, 167–190. https://doi.org/10.1007/978-94-017-0837-1_11
- [40] Charlotte S. Vlek, Henry Prakken, Silja Renooij, and Bart Verheij. 2016. A method for explaining Bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law* 24, 3 (Sept. 2016), 285–324. <https://doi.org/10.1007/s10506-016-9183-4>
- [41] C. Yvan, H. Lim, and T. Lu. 2011. Most Relevant Explanation in Bayesian Networks. *Journal of Artificial Intelligence Research* 42 (Nov. 2011), 309–352. <https://doi.org/10.1613/jair.3301>
- [42] Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. <https://doi.org/10.48550/arXiv.1311.2901> arXiv:1311.2901 [cs].

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009