CausalChat: Interactive Causal Model Development and Refinement Using Large Language Models

Yanming Zhang, Akshith Kota, Eric Papenhausen, and Klaus Mueller, *Fellow, IEEE*

***Abstract*—Causal networks are widely used in many fields to model the complex relationships between variables. A recent approach has sought to construct causal networks by leveraging the wisdom of crowds through the collective participation of humans. While this can yield detailed causal networks that model the underlying phenomena quite well, it requires a large number of individuals with domain understanding. We adopt a different approach: leveraging the causal knowledge that large language models, such as OpenAI's GPT-4, have learned by ingesting massive amounts of literature. Within a dedicated visual analytics interface, called CausalChat, users explore single variables or variable pairs recursively to identify causal relations, latent variables, confounders, and mediators, constructing detailed causal networks through conversation. Each probing interaction is translated into a tailored GPT-4 prompt and the response is conveyed through visual representations which are linked to the generated text for explanations. We demonstrate the functionality of CausalChat across diverse data contexts and conduct user studies involving both domain experts and laypersons.**

***Index Terms*—Human computer interaction (HCI), Explainable AI, Large Language Models, Visualization.**

I. Introduction

CAUSAL relationships are the building blocks of how we make sense of the world. They help us understand why things happen the way they do, from the simple cause-and-effect of a light switch to the complex interplay of factors influencing societal trends. We encounter causal facts in everyday life, whether in conversations about health choices or discussions on broader issues like health policies.

Beneath this universal concept lies a long-standing philosophical debate between Hume and Kant. Hume, an empiricist, saw causality as an expectation formed through repeated experience [1], while Kant, a transcendental idealist, regarded it as an inherent mental framework that shapes how we perceive the world [2]. This tension echoes in modern causal inference: data-driven methods, like Hume's, rely on statistical patterns, while text-based methods, in Kantian spirit, extract causal knowledge from language or judgment. Both have limitations—data-driven methods struggle with confounders and overfitting [3], while text-based ones face ambiguity and difficulty distinguishing causation from correlation [4].

In this paper, we integrate perspectives from Kant and Hume in advancing causal analysis. To incorporate Kant's perspective we leverage large language models (LLMs) trained on extensive text data, and so harness their rich causal knowledge to minimize ambiguities. Following Hume's approach, we also incorporate data when available. While this combined strategy is not entirely new [5], to our knowledge this approach has not been explored thus far with LLMs. It allows users to navigate and aggregate complex causal relationships, enhancing accessibility even for those without domain expertise.

A widely recognized issue with LLMs, like GPT-4, is their tendency to produce inaccurate information or hallucinations, even with well-constructed prompts [6]. To address this, we explore causal questions from multiple directions and polarities of the relationship. This strategy helps reveal the broader context, at GPT-4 often mentions latent, confounding, and mediating variables—offering valuable insights. Additionally, users can also explicitly request these variables, and so substantially expand the scope of the causal model. However, the richness of these multifaceted prompts comes at a cost: they generate substantial textual output that can overwhelm users, even with summarization. To support analysts in navigating and interpreting this complexity, we introduce interactive visualizations that distill and organize the generated insights.

This human–LLM collaboration reflects a clear division of labor: the LLM surfaces plausible causal hypotheses—such as links, mediators, or confounders—while the human evaluates and selects from these alternatives based on context, experience, or further inquiry. Rather than prescribing a fixed workflow, the system supports flexible, user-driven exploration, accommodating variability in how users engage with causal reasoning. This conversational dynamic—where causal knowledge is elicited, examined, and shaped through dialogue between user and model—inspired the name *CausalChat*.

In summary, our contributions are as follows:

- A human-in-the-loop framework for developing and refining causal networks using LLMs and data, embodied in a conversational workflow where users iteratively probe variables, review hypotheses, and curate causal structures.
- A prompt design strategy that interrogates hypothesized causal relationships from diverse perspectives, including mediators, confounders, and latent variables.
- An interactive visual interface for recursive exploration and model refinement, featuring dedicated visualizations for multi-perspective LLM-generated explanations.

Fig. 1 shows our visual analytics dashboard for a causal model of a car. In the following we describe related work (Section II), our methodology (Section III), some usage scenarios (Section IV), a user study (Section V), a discussion (Section VI), and conclusions (Section VII).

All authors are with the Computer Science Department at Stony Brook University. Contact author: yanming.zhang@stonybrook.edu

Manuscript received September 19, 2024; revised 2025, accepted 2025

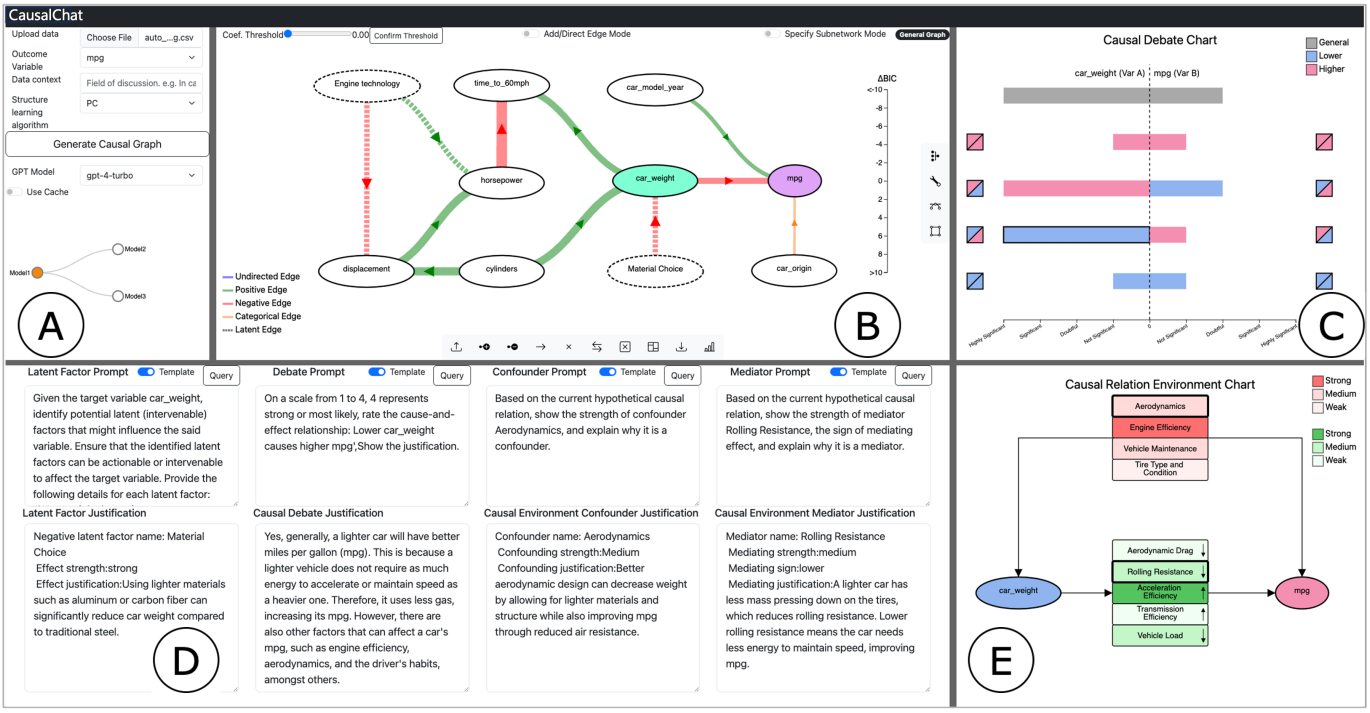

Fig. 1. The CausalChat Dashboard analyzing the AutoMPG dataset. It includes (A) the Control Panel allowing users to specify fundamental parameters and manage the model tree for variations of the causal model, (B) the Causal Graph Panel for interactive refinement of the graphical model, (C) the Causal Debate Chart for resolving causal directions and inclinations, (D) the Causal Justification Panel offering a rationale for each hypothetical causal statement, covering latent factors, potential confounders, and mediators, and (E) the Causal Relation Environment Chart suggesting potential latent variables, confounders, and mediators for specific causal relations and variables. In this figure the confounder/mediator chart is shown.

II. Related Work

Causal networks are widely used in various fields, such as epidemiology [7], healthcare [8], biology [9], and social sciences [10] to understand complex systems. In observational studies, causal networks are typically represented using directed acyclic graphs (DAGs), which clarify causal structure and help identify sources of bias, such as confounding and selection effects. Constructing such a DAG is best achieved through the rigorous process of randomized controlled trials based on first principles. However, this method often faces challenges, be it due to cost, ethical considerations, or practical limitations. It is also not easily scalable and may restrict the number of researchers who can engage in such studies. A more scalable and general approach involves deriving a causal DAG through one of three primary methods: analyzing data, analyzing text, or collaborative construction and crowdsourcing. Many of these approaches combine elements from more than one of these paradigms, and some allow human analysts to participate in the DAG development process [11]–[13].

A. Causal Network Discovery using Numerical Data

There are essentially two popular strategies for causal discovery. One approach involves enforcing the constraint that two statistically independent variables are not causally linked, followed by a series of conditional independence tests to construct a compliant DAG. Well-known algorithms for this method include the PC algorithm [14] and the Fast Causal Inference (FCI) algorithm [15]. Another strategy is to greedily explore the space of possible DAGs via Greedy Equivalence Search (GES) [16]. This entails score-based methods in which edges are iteratively added and removed from the graph to maximize a model fitness measure, such as the Bayesian Information Criterion (BIC) [17], [18]. Then, once the causal structure has been learned, Structural Equation Modeling (SEM) [19] is applied to estimate the strength and sign of the causal effects, typically by fitting linear models.

Causal discovery relies on four common assumptions: (1) the causal structure can be represented by a DAG, (2) all nodes are conditionally independent of their non-descendants given their parents, (3) the DAG is faithful to the underlying conditional independence, and (4) the DAG is sufficient, i.e., there is no pair of nodes with a common external cause. Unfortunately, these conditions are rarely entirely met, often due to selection/sampling bias in the data. Essentially, the phenomenon to be explained by the causal network is only partially captured by (1) the measured variables and (2) the observed data samples, and this leads the discovery algorithm astray. While the probability of obtaining a partially incorrect DAG can be reduced by using more data, it remains uncertain how much data is truly needed [20], [21].

Our research tackles both of these bottlenecks: (1) the limitation of collected datasets to fully capture all variables essential for a comprehensive causal model – we address this through GPT-4-based variable ideation and relevance assessment, and (2) the absence of data for the newly discovered links – we use GPT-4 to generate plausible estimates based on its contextual knowledge and learned patterns.

B. Causal Network Discovery using Textual Data

While extracting causal relations from text documents is not a new endeavor [22], thanks to ChatGPT this process has become remarkably convenient with a simple prompt. The literature on LLM assisted causal network learning is rapidly expanding. The earliest documented attempt using LLMs (specifically GPT-3) for causal analytics was by Long et al. [23]. However, this was a preliminary study focused on optimizing prompts to reveal insights into the presence or absence of directed edges. More recently, Kıcıman et al. [24] delved much deeper into the subject. They devised a comprehensive set of prompts for GPT-3.5 and GPT-4, generating yes/no responses to standard causal queries. While they demonstrated excellent success rates on benchmark datasets where causal truth was known, they did not explore utilizing GPT's output to gather additional causal and contextual knowledge. Similarly, subsequent papers (e.g., [25]–[27]) also did not explore visualizing the acquired information within explainable and trustworthy AI.

Some studies have highlighted the limitations of using LLMs for causality analysis. While LLMs excel at discerning causality from empirical or commonsense knowledge, Jin et al. [28] demonstrated their significantly reduced effectiveness in deriving causality through pure causal reasoning—something numerical algorithms like PC are specifically designed for. To test this, they assembled a substantial dataset comprising over 400,000 correlational statements in natural language and tasked the LLM with determining the causal relationship between variables. Their findings revealed that existing LLMs demonstrated performance is akin to random chance in this particular task. These findings are echoed by Zečević et al. [29], who suggest that LLMs can serve as a valuable starting point for learning and inference, reaffirming their role as a tool for ideation and creativity. They can complement data-driven causal inference methods, such as PC, which is what one of the approaches we promote here in this paper.

C. Collaborative Causal Network Discovery with Crowds

In 2018, Berenberg and Bagrow [30] introduced a methodology that harnessed the 'wisdom of the crowds' to construct a large causal network, utilizing the widely-used crowdsourcing platform Amazon Mechanical Turk. They devised a three-stage approach: in stage 1, workers proposed causes; in stage 2, they suggested effects for these causes; and in stage 3, they edited and refined longer causal pathways derived from the stage 2 results. The final causal network was then formed by amalgamating all worker-generated pathways, with more popular edges indicating stronger causal links. Salim et al. [31] adopted a similar approach, focusing on mining crowd beliefs and misconceptions in complex systems with societal impact such as climate change.

It is worth noting that the study by Berenberg and Bagrow predates the emergence of LLMs. While the degree to which LLMs, trained on extensive human-written text, tap into the 'wisdom of the crowds' remains uncertain, it is plausible to expect that LLM assistance would necessitate a significantly smaller crowd. As LLMs effectively encapsulate the viewpoints of a large crowd simultaneously, using a few-shot prompting approach can guide and constrain the response towards a pertinent answer [32]. We believe that our multifaceted prompt represents a significant step in this direction.

Yen et al. [33] developed an interactive system for collaborative causal network construction. This system enabled users to articulate narratives to explain causal relationships they perceived, visualize the causal models of these using DAGs, and review and incorporate the causal diagrams and narratives of other users. A notable feature of their system was the 'Inspire Me' popup, which users could request when they needed fresh ideas on how to expand the network. They would then be presented with one of several pre-programmed thought-provoking questions related to causal relationships.

The purpose of this system was to investigate whether actively evolving and narrating a causal network, and learning from networks constructed by peers, could uncover blind spots in a person's causal reasoning and lead to a refinement of their own causal network. More recently, the authors introduced CrowdIDEA [5], an enhanced interface that includes a data panel with visualizations and statistics. It is conceivable that an LLM could fulfill a similar collaborative role. For instance, in our system users can explore any variable or pair and visualize suggested directions, confounders, and mediators—ready for seamless integration into the emerging causal network

D. Causal Network Discovery with Visual Analytics

Visual analytics bridges the gap between machine learning outputs and human goals, enabling users to guide, interpret, and refine results—especially when models are imperfect or lack a full understanding of the task. Most existing visual causal analysis tools were developed before the widespread use of LLMs and thus focus on numerical data. These systems often enable human-in-the-loop causal refinement. DAGitty [34] was among the first to enable interactive DAG creation and analysis using graphical representations alone. Wang et al. [11] introduced one of the earliest visual interfaces for interactive causal reasoning with causal networks and "what-if" simulations, powered by algorithms for causal link discovery and statistical evaluation, further extended to identify subgroup-specific causal networks in heterogeneous datasets [12] and temporal data [35]. More recently, Guo et al. [36] presented Causalvis, a Python toolkit supporting DAG creation, subgroup matching, and outcome estimation with visualizations.

In this paper, we introduce a new visual framework that integrates data-driven discovery, LLM-generated hypotheses, and human oversight. The interface coordinates these components, allowing each to mitigate the others' limitations. This builds on preliminary ideas from [37], now extended with a refined implementation and improved design.

III. Methodology

Our workflow is depicted in Fig. 2. It begins with tabular data, which is processed by a Structure Learning Algorithm – we use GES – to derive its Causal Structure, capturing the potential causal relationships among the variables. Simultaneously, Structural Equation Modeling (SEM) is used to

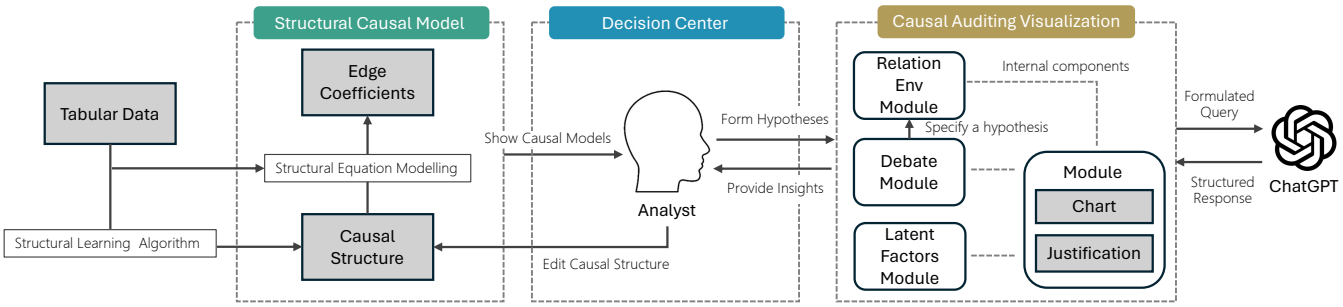

Fig. 2. Workflow of our ChatGPT-powered causal graph development environment. Starting from an incomplete causal model—either algorithmically generated or manually built—the model is iteratively expanded and refined by a human expert via a GPT-4-powered conversational visual interface.

estimate Edge Coefficients that quantify the strength of these relationships. Together, these elements form the Structural Causal Model.

This model is then presented in the Decision Center, where an analyst (1) explores the causal relationships, (2) forms hypotheses, and (3) refines the model by editing the causal structure as needed. The analyst plays a central role in interpreting and shaping the model. To support this process, the system includes a Causal Auditing and Visualization environment, which is subject of this paper. This environment consists of three key modules: (1) the Relation Environment Module, which lets users specify hypotheses about possible causal variables and links; (2) the Debate Module, which compares competing causal explanations; and (3) the Latent Factors Module, which identifies hidden or unobserved variables that might explain observed relationships.

Each module provides output through two integrated components: (1) a set of Charts that visualize numerical evaluations such as causal strength scores, and (2) a Justification Panel that offers textual explanations. These components are powered by ChatGPT, which receives a formulated query—automatically generated based on the selected hypothesis and context—and returns a structured response containing both the score and the rationale. Together, these elements enable a human-in-the-loop workflow where visual and textual insights work in tandem to support informed, transparent, and iterative causal modeling. In the process. users may locate additional tabular data to support SEM-based link parameterization or they might map the edge scores provided by GPT-4 into edge weights.

A. Prompting for Direct Causal Relationships

The prompts[1] we utilize follow an optimized template (see Appendix B.1 and B.4). Offering sufficient contextual information and guidance on expectations is crucial for prompt engineering [32], [39]. Below is an abstraction of the prompt.

Prompt: You are an expert in <domain>. On a scale from 1 to 4, where 4 represents highly significant, 3 represents significant, 2 represents doubtful, 1 represents not significant, rate the following cause-and-effect relationship: Does higher/lower A/B cause higher/lower B/A.

[1]Due to the inherent stochasticity of LLMs, the same prompt may yield different outputs across runs. To improve reliability, we adopt a self-consistency approach [38], running each scoring prompt 10 times and selecting the most frequently returned score. To minimize token usage, we request a justification only for this majority score. Appendix B.4 shows some score histograms.

This generates 10 distinct prompts, 5 each for A and B taking opposite roles, and within each of these two sets there are 4 combinations of A and B being (higher, lower) plus one relation that just asks this for a general case. An example prompt is shown below, where (...) denotes further prompt specifications (see Appendix B.1.1 for complete prompt):

Prompt: You are an expert in public health. On a scale from 1 to 4, where 4 represents highly significant, 3 represents significant, 2 represents doubtful, 1 represents not significant, rate the cause-and-effect relationship: Does *higher percent fair or poor health* cause *lower life expectancy*...

Including the domain hint 'public health' provides contextual information. GPT-4 can also infer the domain from the dataset attributes if it is told to do so. GPT-4's response to the prompt is to the point:

Response: Rating: 4

B. Prompting for Confounders

This prompt template (see Appendix B.2.1) also distinguishes among the 4 combinations that explore the effects of higher and lower levels plus one relation that just asks this for the general case. In the following we explain this template using the variables *food environment index* and *violent crime rate* as an example. Also here (....) denotes omissions.

Prompt: You are an expert in public health. Given the cause-and-effect relationship 'lower food environment index' causes 'higher violent crime rate' identify potential confounders based on the definition .... For each identified confounder, provide the following details in a tuple format: 1. Name of the confounder. 2. Strength of the confounder (options: weak, medium, strong). 3. Justification for its role as a confounder based on the definition provided.

Response: GPT-4 returned 2 'strong' confounders (Socioeconomic Status, Residential Segregation) and 4 'medium' confounders (Substance Abuse and Mental Health Issues, Availability of Public Services, Racial and Ethnic Composition, Neighborhood Disorganization). For each a detailed justification was given, such as "Substance abuse and mental health issues can contribute to both a lower food environment index (due to prioritization of immediate needs over healthy food choices) and higher rates of violent crime, as these issues can lead to unstable social environments".

C. Prompting for Mediators

Also here we distinguished among the 4 level combinations and the general one. Using the same example as for the confounder, the (partial) mediator prompt is below (see Appendix B.2.2 for complete prompt).

Prompt: You are an expert in public health. Given the cause-and-effect relationship 'lower food environment index' causes 'higher violent crime rate' identify potential mediators based on the definition: Rather than a direct causal relationship between the independent variable and the dependent variable, the independent variable influences the mediator variable, which in turn influences the dependent variable. For each identified mediator, provide the following details in a tuple format: 1. Name of the mediator. 2. Strength of the mediator (options: weak, medium, strong). 3. Justification for its role as a mediator (...) 4. Specific conditions under which the mediator operates (...) 5. Direction of the mediator's effect ('positive' or 'negative') (...). The direction tells us how to intervene on the mediators to achieve the relationship....

The 'Direction' parameter (parameter 5) is crucial as it specifies the way in which the level of a mediator should change to influence the effect variable as indicated. This guidance not only informs analysts about the type of intervention required but also sheds light on the underlying cause-effect mechanism.

Response: GPT-4 returned 1 'strong' mediator (Economic Disadvantage ↑) and 4 'medium' mediators (Social Cohesion ↓, Substance Abuse ↑, Educational Attainment ↓, Mental Health ↓), where ↓ ↑ indicate the direction the mediator needs to have to support the effect. For each mediator a detailed justification was given, such as "A lower food environment index may contribute to reduced social cohesion within a community, as limited access to nutritious food options can lead to increased stress and poorer overall health. Reduced social cohesion has been associated with higher rates of violent crime, as it may lead to weaker community bonds and less effective informal social control".

D. Prompting for Latent Factors

Unlike previous prompts, this prompt focuses on a single variable, emphasizing intervenable factors. It identifies actionable variables as points of intervention, enabling practitioners to influence the target variable through specific causal pathways. An example of a latent factors prompt is provided below (see Appendix B.3 for the full prompt).

Prompt: Given the target variable *primary care physicians rate*, identify potential latent (intervenable) factors that might influence the target variable. Ensure that the identified latent factors can be actionable or intervenable to affect the target variable. Provide the following details for each latent factor: 1. Name of the latent factor. 2. Strength of the effect (weak, medium, strong). 3. Sign of the effect (positive, negative, or categorical). 4. Justification for its role as a latent factor.

Response: GPT-4 returned 1 'strong' positive latent factor (Reimbursement Rates), 2 'medium' positive latent factors (Medical Infrastructure Investment and Healthcare Policy Reforms), 1 'strong' negative latent factor (Medical Student Debt), and 1 'medium' negative latent factor (Urbanization Incentives). For each latent factor a storyline rationale was given. For instance, Medical Student Debt can serve as a negative latent factor, as 'high levels of debt from medical education can deter graduates from entering lower-paying specialties like primary care.' Medical Student Debt can be addressed through governmental medical debt relief programs, which act as intervention points.

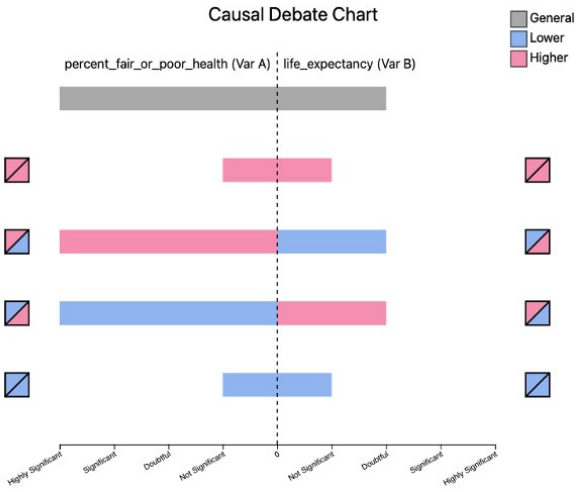

Fig. 3. Causal Debate Chart for the relation *Percent Fair or Poor Health - Life Expectancy*, presenting an overwhelming belief that the former is the cause of the latter.

E. Visualizing the GPT-4 Generated Text Responses

While the inclusion of text helps justify the presence (or absence) of a causal relation, GPT-4 may generate excessive text. Even when instructed to summarize its findings, this abundance of information can overwhelm general users. Below, we present the visualizations we have designed to make browsing this information easier.

F. The Causal Debate Chart

The Causal Debate Chart summarizes the distribution of GPT-4-generated strength scores for the 10 causal prompts introduced above, revealing both directional support and variability in the responses. By aligning scores with increasing or decreasing values of the causal variable and its directionality, the chart enables users to visually assess the plausibility and stability of a given causal assertion. We call it the Causal Debate Chart because it visually argues the strength of one variable being the cause of the other—much like a "debate."

Fig. 3 shows an example of this chart, a bidirectional bar chart [2] where each side is headed by one of the two relation variables. In this case the left side is *Percent Fair or Poor Health (PFPH)* and the right side is *Life Expectancy (LE)*. The x-axis is the score assigned by GPT-4 and the length of

[2]While we use bar lengths for visual clarity, they function as level indicators of the causal strengths reported by GPT-4—not as precise numeric values.

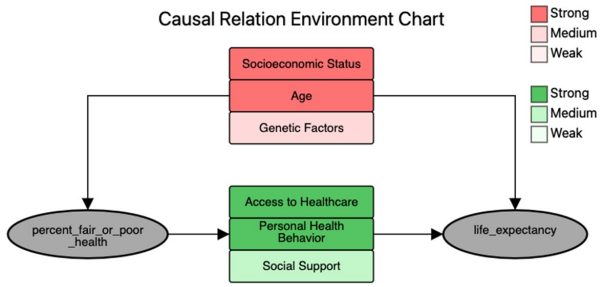

Fig. 4. Causal Relation Environment Chart for the relation *Percent Fair or Poor Health - Life Expectancy*. The intensity of red and green encodes the strength of the mediators and covariates (weak, medium, strong), and the color of the cause and effect variables have the same interpretation as those in Fig. 3; in this specific case they are grey.

each bar is mapped to that score. The grey bars are for the general prompt while the other bars are colored in magenta if the cause was a higher or increasing level of the variable or in sky blue if the cause was a lower or decreasing level (see color legend on the top right). The textual level descriptions have been deliberately chosen to be relatable to humans.

Let us now evaluate the chart. We observe that for the first (grey) set of bars *PFPH* has a substantially longer bar (level 4) than *LE* which has level 2 (level 2 is a doubtful cause in GPT-4 semantics). It means that the former wins the causal debate – it has causal dominance. *PFPH* seems to be a general cause of *LE*.

Let's examine the other bars representing specific level studies. Here, we assess whether GPT-4 maintains consistent logic (as opposed to hallucinating). We observe that high or increasing *PFPH* leads to low or decreasing *LE* in the third set, and the same holds for the opposite relation in the fourth set. Sets two and five display low bars on both sides, as expected if the relation indicated by the other bars is considered true. The Causal Debate Chart in Fig. 3 serves as a prime example of what we would expect from a steadfast causal relation.

G. The Causal Relation Environment Chart

The Causal Relation Environment Chart supports structural reasoning by showing not only the direct relationship between two variables, but also the surrounding context: mediators, confounders, and latent variables identified through LLM queries. It provides a higher-level view of the causal system, helping users consider indirect pathways, common causes, and alternative explanations. Fig. 4 shows an example of this chart for the general *Percent Fair or Poor Health (PFPH) - Life Expectancy (LE)* relation, where confounders are colored in shades of red and mediators n shades of green – the shading reflects their strength (see color legend on the right).

It is often the case that GPT-4 will identify similar mediators and confounders for the more focused (low, high) relations. But they vary in the sign. For example, to go from *low PFPH* to *high LE*, positive levels of the mediators are cited, such as good access to healthcare and good health habits, while to go from *high PFPH* to *low LE* the cited mediators are usually the opposite, like limited access to health care and poor health habits. Fig. 5 shows these two cases where the up and down arrows indicate the positive and negative levels, respectively.

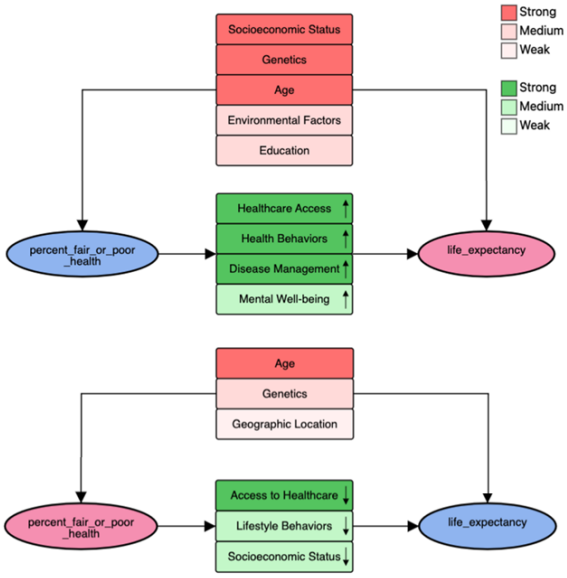

Fig. 5. Causal Relation Environment Chart for two level combinations of the relation *Percent Fair or Poor Health - Life Expectancy*. The up and down arrows show the appropriate signs of the mediators.

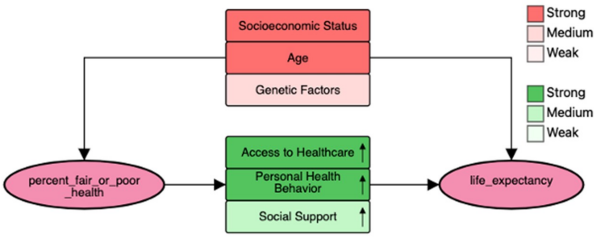

Fig. 6. Causal Relation Environment Chart for an improbable level combination of the relation *Percent Fair or Poor Health - Life Expectancy*, namely one where both variables have positive levels. The up arrows in the mediators show how this improbable combination might be achieved, in form of interventions on the mediators in the direction of the arrows.

Fig. 6 explores the unlikely relationship of *high PFPH* causing *high LE*. GPT-4 correctly identifies this as "counter-intuitive" but treats it as a 'hypothetical scenario'. It suggests that a mediating relationship would need to exist to achieve the desired *high LE*. These mediators represent potential points of intervention that policymakers could target to increase *high LE*, despite the *high PFPH*. For example, a policymaker might intervene by opening additional health clinics in areas with *high PFPH*, thereby increasing the mediator, *Access to Healthcare*, which in turn improves *high LE*.

H. The Latent Factors Chart

The Latent Factors Chart is designed to advise on supplementary variables that should be included in the analysis of the current causal graph. It highlights the variable of interest alongside other potential factors that may directly influence it. These latent factors are color-coded: different shades of yellow represent negative effects, while different shades of blue represent positive effects. By reviewing the provided justifications, users can verify the relevance of latent factors and incorporate the most significant into the causal graph—a process supported by prior work [40] showing that AI-generated explanations enhance user trust and confidence.

For example, as shown in Fig. 7, *Reimbursement Rates* is identified as a strong positive influence on the *Primary Care Physician Rate*. GPT-4 justifies this by stating, 'Higher reimbursement rates for primary care services can make the field more financially appealing, attracting more physicians to primary care and directly influencing the primary care physician rate.' On the other hand, as discussed in section III-D, *Medical Student Debt* is a strong negative factor, with high levels of debt deterring graduates from entering lower-paying specialties like primary care. To improve the primary care physician rate, a policy analyst may update the causal graph to emphasize increasing reimbursement rates and reducing medical student debt as key intervention points.

I. Model Tree

Although DAGs provide intuitive representations for causal graphs, their scalability is limited, with typical graphs containing an average of 12 nodes, with most graphs having between 9 and 16 nodes [41]. To address the challenge of representing causal relationships involving dozens or even hundreds of variables while still using DAGs, we introduce the Model Tree (see Fig. 1(A)).

The Model Tree is an N-ary tree structure in which each node represents a distinct causal graph. The root node corresponds to the global model, providing an overview of the entire system, while child nodes represent progressively more specific or local models. Users interact with the general model at the root and can select a subset of nodes to create a child node in the Model Tree. These child nodes inherit a subgraph from the parent node, composed of the selected nodes and their associated edges. As users continue to refine and expand the causal graph at each hierarchical level, the Model Tree evolves into a structured hierarchy of causal graphs, offering different levels of granularity and detail.

The Model Tree also supports personalization of causal models. Bidirectional effects, or feedback loops, occur when two variables influence each other over time [42]. A classic example is the relationship between obesity and depression, where each can be a cause of the other over time. While such bidirectional effects are prevalent in fields such as epidemiology, biology, etc., they conflict with the acyclic nature of DAGs. The Model Tree bypasses this issue by splitting a causal model with a bidirectional edge into two causal models with unidirectional edges, thereby maintaining acyclicity. This approach aligns with the concept of personalized causality or causal heterogeneity, where the direction of causal relationships may differ between individuals or for the same individual at different points in time.

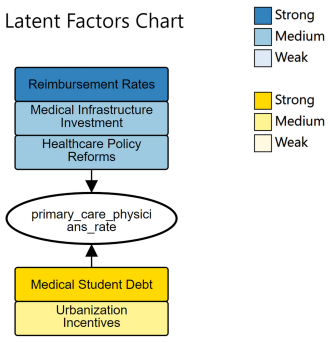

Fig. 7. Latent Factors Chart for *Primary Care Physicians Rate* as the target variable. The blue nodes above are factors with positive influence, while the yellow nodes below are factors with negative influence, intensity codes strength (see color legend on the right).

J. The Causal Justification Panel

The Causal Justification Panel plays a key role in causal model development by providing GPT-4–generated natural language explanations for each suggested causal relationship, grounded in general knowledge. It complements the numerical strength scores shown in the charts by adding context and interpretability, helping users understand why a link or variable might be considered strong or weak. While the visualizations support streamlined exploration and comparison, the Justification Panel offers a vital textual counterpart that enables users to validate, question, or refine their hypotheses. Together, these components foster more informed and explainable causal modeling—especially for users without deep domain expertise.

The panel consists of four sub-panels, each corresponding to one of the charts (see Fig. 1D, from left to right): latent factors, debate relations, confounders, and mediators.They are filled on demand: when a user selects an element from one of these charts, the corresponding sub-panel is populated with the GPT-4 prompt automatically generated from a template, followed by the resulting score and its justification. Users can also write their own prompt by de-selecting the 'Template' switch and clicking 'Query', enabling a conversation with GPT-4.

K. Graphical Encoding Schemes and Dashboard

In the graphical encoding schemes, we sought to maintain clarity across different diagrams while ensuring consistent meaning for shapes and colors that represent similar concepts. As mentioned, in the causal diagram, red lines (varying in thickness) represent positive causal relationships, and green lines represent negative causal relationships. The variation in thickness reflects the strength of these relationships. Different shapes provide context: oval-shaped purple marks the outcome variable, cyan denotes a selected node, unfilled ovals signify causal nodes, and dotted shapes/lines indicate elements derived from GPT-4 that are not yet confirmed by data.

In the Causal Debate Chart, using magenta bars for increasing levels and sky blue bars for decreasing levels provides a clear visual contrast, ensuring that this chart remains visually distinct from the others, especially in its portrayal of variable levels. The Causal Environment Chart uses red shaded boxes for confounders and green shaded boxes for mediators, maintaining consistency with the causal diagram's color scheme but applying the colors to shaded boxes (instead of lines), which helps differentiate between the two types of charts.

Finally, the Latent Factors Chart introduces blue shaded boxes to represent positive influence and yellow shaded boxes for negative influence. These colors were chosen to avoid overlap with the red and green used in other charts, allowing the latent chart to stand apart while still adhering to a recognizable positive/negative color scheme.

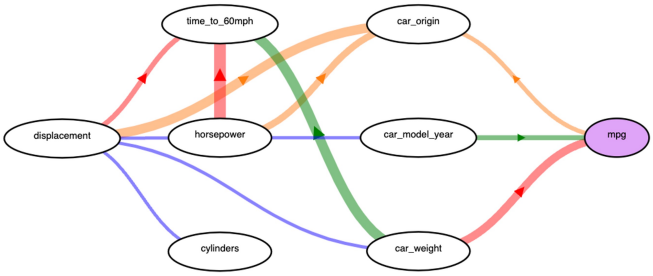

Fig. 8. Initial causal graph of the AutoMPG dataset generated by the GES algorithm. Green (red) directed edges indicate positive (negative) causation, blue undirected edges link variables that are correlated and potentially causal, yellow causal edges connect categorical variables.

All charts and GPT-4 justifications are accessible via an interactive dashboard that implements these graphical encodings. Fig. 1 shows an example of this dashboard for the first usage scenario, presented next.

L. Implementation Details

CausalChat uses a Python backend to manage GPT-4 API interactions and integrates data-driven causal inference methods through the causal-learn library for structure learning and the dowhy library for treatment effect estimation. The frontend is built with D3.js to support interactive visualizations.

To generate the Causal Debate Chart, we use chain-of-thought prompting [43], decomposing complex queries into intermediate steps for more consistent ratings. Prompts include rating instructions, causal hypotheses, and response formatting. To reduce latency and avoid redundant calls, GPT-4 responses are cached in BigQuery; for each hypothesis, GPT-4 is queried ten times, and the mode rating is stored.

IV. Usage Scenarios

In this section, we demonstrate the capabilities of CausalChat by presenting two usage scenarios that employ real-world datasets.

The AutoMPG Dataset [44] covers 398 cars from the 1980s. Each car is characterized by 8 attributes: origin, model year, weight, horsepower, displacement, acceleration (time to 60 mph), cylinders, and miles per gallon (mpg). Due to their simple mechanics, 1980s cars exhibit straightforward causal relationships among the variables.

The Opioid Death Dataset combines 9 key socioeconomic factors sourced from the County Health Ranking database [45] with opioid death data from the CDC WONDER database [46] for over 3,000 US counties. The chosen factors are hypothesized to have either direct or indirect impacts on opioid-related deaths.

A. Exploratory Causal Analysis: Automotive Engineering

In this study, we focus on Oscar, an automotive hobbyist who wants to gain more insight into automotive engineering. Oscar finds the AutoMPG dataset and reads it into CausalChat. He then employs the GES algorithm and obtains the preliminary causal graph shown in Fig. 8 (see caption for an explanation of the edge coloring). We now follow Oscar in his mission to audit and expand this causal graph.

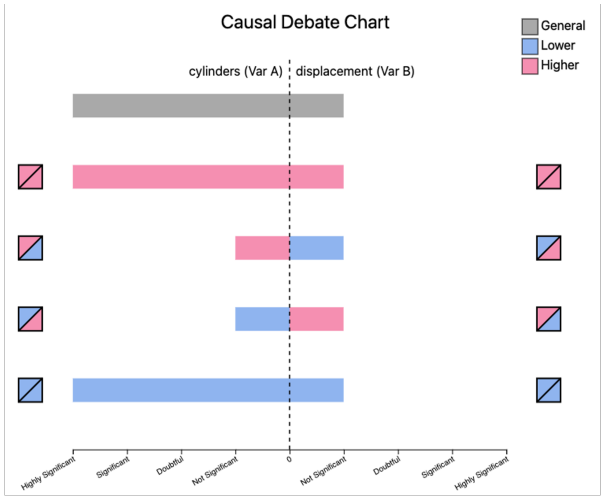

Fig. 9. Causal Debate Chart of the relation *Cylinders - Displacement*.

Resolving undirected edges. Oscar identifies four blue edges that the GES algorithm could not resolve, possibly due to the dataset's limited coverage of the car domain. He employs CausalChat's GPT-4 suite to address these edges. For brevity, we shall focus on how he resolves the blue edge between *Cylinders* and *Displacement*, the procedure for the other blue edges is similar.

Oscar clicks on the blue edge and the system generates the Causal Debate Chart depicted in Fig. 9. It is immediately apparent that the bars representing *Cylinders* are notably longer than those for *Displacement*, for the general grey bars as well as for the red-red and the blue-blue bars and all at full strength – a classic pattern for a positive causation. Oscar contemplates converting the blue edge to a green directed link from *Cylinders* to *Displacement*. After confirming this direction from the GPT-4 justification panel, 2nd column (see an example in Fig. 1D for the relation lower *Car Weight* $\rightarrow$ higher *MPG*), Oscar directs the edge as suggested. In a similar fashion he also directs the other blue undirected edges.

Adding confounders and latent variables. Oscar is still curious about the relation of *Displacement* and *Horsepower*. He examines the corresponding Causal Relation Environment Chart (see Appendix C.2) and identifies *Engine Technology* as a confounder. Furthermore, he also discovers a latent factor – *Material Choice* – which exerts a reducing effect on *Car Weight* (see Appendix C.3). All of these interactions taken together give rise to the final causal graph shown in Fig. 1B. The edges for the two newly added variables are visualized as dotted lines to convey that their weights have not been calculated from data yet – all Oscar has are the strengths indicated by GPT-4.

Inspecting the triad *Displacement*, *Horsepower*, and *Engine Technology* reveals that *Horsepower* is, in fact, a collider[3]. It is influenced by both *Displacement* and *Engine Technology*, or dominated by one of the two. While in older cars displacement

[3]A collider is a variable that is influenced by two or more other variables (e.g., A $\rightarrow$ C $\leftarrow$ B). While the causes (A and B) may be independent, learning about one can change beliefs about the other when the collider (C) is known.

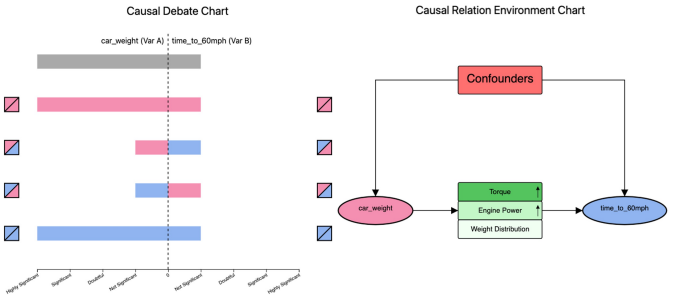

Fig. 10. To assess the hypothetical causal relation where a low time to 60 MPH could be achieved despite high car weight, we inspect the third bar pair in the Debate Chart (left) in the corresponding Environment Chart (right).

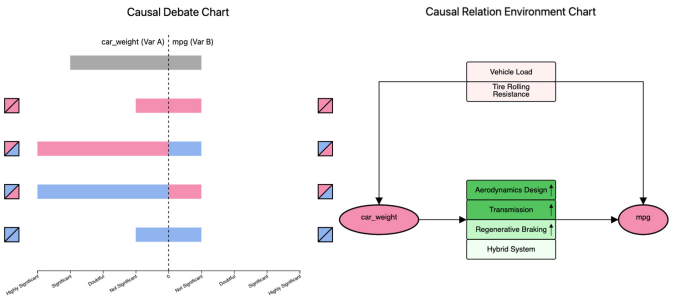

Fig. 11. To assess the hypothetical causal relation where a high MPG could be achieved despite high car weight, we inspect the second bar pair in the Debate Chart (left) in the corresponding Environment Chart (right).

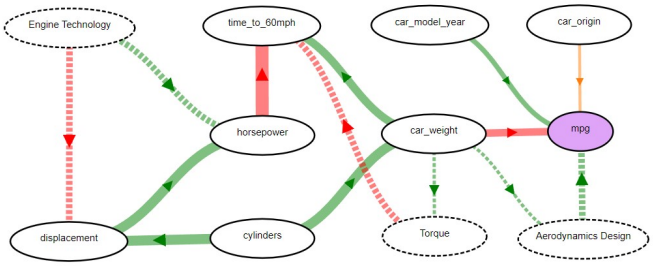

Fig. 12. Causal graph that incorporates all of Oscar's innovations.

was the dominant factor determining horsepower, modern engine technology has altered this role. The negative causal effect of *Engine Technology* on *Displacement* further suggests that as engine technology advances, the need for displacement to elevate horsepower diminishes. In other words, higher values of engine technology have become more influential in determining horsepower than displacement alone. This interpretation aligns with the idea that modern engines, with advancements in technology, can achieve higher horsepower despite low displacement. In essence, by adding *Engine Technology*, Oscar has modernized the original causal network derived from the antiquated dataset of 1980s cars.

Adding mediators. In the updated causal graph Oscar notices the antagonistic relationship of *Car Weight* and *Horsepower* which affect *Time to 60 MPH* in opposite ways. Having resolved the need for high *Displacement* also reduces the need for a high number of *Cylinders* which would cause high *Car Weight*. However, there may be other valid causes for high *Car Weight* not represented in the graph. Consequently, Oscar chooses to investigate potential ways to mitigate these factors.

The Causal Debate Chart in Fig. 10 (left) illustrates that heavy cars take more time to reach 60 MPH (2nd bar pair) than lighter cars (5th bar pair), i.e., they have poor acceleration. These are the hard facts represented by the two long bars. But Oscar wants to innovate and is looking for a car that can be heavy yet quick off the mark, the condition represented by the 3rd pair of bars. Clicking on the left bar brings up the corresponding Causal Relation Environment Chart, Fig. 10 (right). It suggests that increasing the engine's twisting power, or torque, can improve acceleration and mitigate the effect of high car weight.

This prompts Oscar to include *Torque* as a mediator between *Car Weight* and *Time to 60 MPH*, see Fig. 12. It introduces *Torque* as an additional causal factor that opposes the impact of high *Car Weight*. This influence is indicated by the red dotted outgoing edge and its purpose is indicated by the green dotted incoming edge, i.e., heavier cars need more torque.

To further optimize his ideal car, Oscar shifts his attention towards enhancing its mileage efficiency. The causal graph reveals an inverse relation link between *Car Weight* and *MPG*. Using a similar process as above he refers to the Causal Debate and Causal Relation Environment Charts in Fig. 11, now for the case of high *Car Weight* and high *MPG*. He learns that incorporating advanced *Aerodynamic Design* principles in the car's design can markedly cut down fuel consumption during operation, ultimately leading to a notable improvement in *MPG*. It advises car designers to apply aerodynamic principles when a car is heavy, as indicated by the green causal edge. Oscar inserts this mediator into the *Car Weight* and *MPG* relation, yielding the final causal graph in Fig. 12.

In this graph the two added mediators turn both *MPG* and *Time to 60 MPH* into colliders, where the newly introduced variables help mitigate–or even neutralize–the adverse effects linked to the original causal factor, *Car Weight*.

B. Causal Strategizing for Opioid Mortality Prevention

Here, we join Lena, an epidemiologist, on her mission to discover preventive measures against the widespread opioid epidemic afflicting numerous counties in the United States. She starts out with 9 socioeconomic variables that she feels are related to opioid mortality plus data on opioid mortality itself (the aforementioned opioid death dataset).

Initial setup. As a first step, Lena utilizes her expertise to construct a foundational causal graph based on the available data, see Fig. 13a. However, she finds herself dissatisfied with the connection between *Education Index* and *Opioid Dispensing Rate*. Intuitively, she believes that enhanced education would raise awareness of the adverse effects associated with opioid dispensing. Yet, the green edge she initially drew suggests the opposite. To investigate, she opts to reassess this edge using the Causal Debate Chart shown in Fig. 14 (left).

Exploring doubts. Examining this chart, Lena comes to realize that while there is a slight inclination towards the current causality, none of the bars exhibit highly significant strength, indicating a raised likelihood of a confounder. She clicks on the left bar of the general (grey-colored) relation which brings up its Causal Relation Environment Chart. Indeed, two

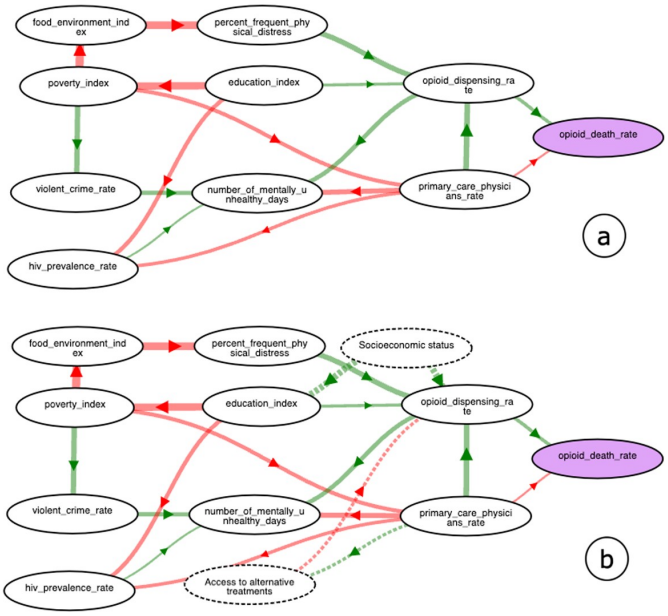

Fig. 13. Causal graphs created by Lena: (a) Initial model leveraging domain knowledge. (b) Model enhanced via CausalChat assistance, incorporating confounders and mediators for actionable recommendations.

confounders are suggested, with *Socioeconomic Status* being the strongest. The justification states that "socioeconomic status can influence level of education and also lead to better access to healthcare facilities where more prescription opioids are dispensed". This confirms Lena's initial apprehension about the direct edge, with the justification pointing to an important mediator between *Socioeconomic Status* and *Opioid Dispensing Rate*: *Prescription*. Thus, a mitigating policy intervention would be to tighten opioid prescription regulations. Finally, Lena also examines the relation between the *Number of Mentally Unhealthy Days* and the *Opioid Dispensing Rate*. This leads to a deep engagement with the Justification Panel which is detailed in Appendix C.1.

Addressing the target effect. Next, Lena sets out to identify actionable measures to help reduce opioid fatalities. She directs her attention to the edge from *Primary Care Physician Rate* to *Opioid Dispensing Rate*. Generally, the relationship is positive since opioid dispensing typically involves doctors[4] as is indicated by the dominant 2nd bar pair in the associated Causal Debate Chart, Fig. 15. In search of an intervention, Lena focuses on the hypothetical, but more desirable relationship just below this pair: higher *Primary Care Physician Rate* leading to lower *Opioid Dispensing Rate*. The associated Causal Relation Environment Chart offers numerous valid mediators, such as *Access to Alternative Treatments*. This suggests that if alternative non-opioid treatments are made available, the opioid dispensing rate and its subsequent use can be reduced. Fig. 13b shows the updated causal graph.

Focusing on a specific population group. Lena now continues her exploration with a focus on a specific causal pathway: *Food Environment Index* → *Percent Frequent Physical Distress* → *Opioid Dispensing Rate* → *Opioid Death Rate*.

[4]Here we consider the original source of opioids: physicians. However, in modern times, opioids often stem from the illicit distribution of fentanyl.

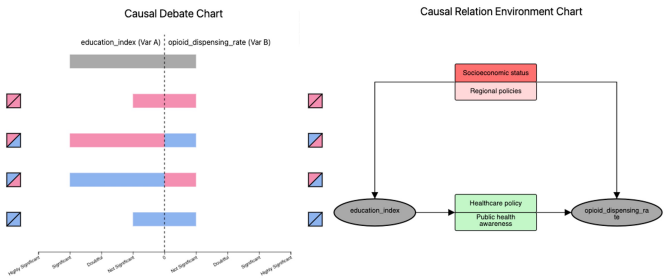

Fig. 14. The weak direct causal relations (left) suggest a confounding between *Education Index* and *Opioid Dispensing Rate* (right).

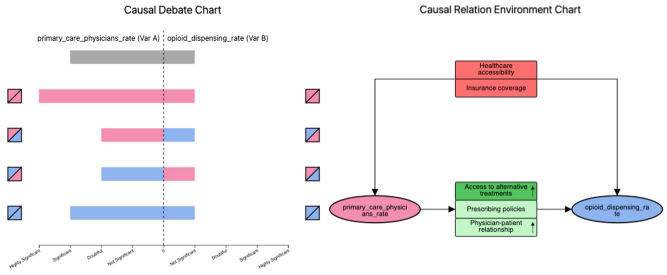

Fig. 15. Gaining control over opioid dispensing through mediators. Shown is the Environment Chart (right) of the third bar pair in the Debate Chart (left).

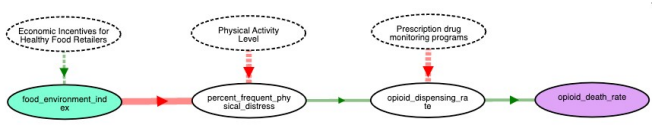

Fig. 16. A causal pathway delineating multiple intervenable latent factors along a critical causal chain for a specific population group.

This pathway tells a compelling story: individuals experiencing a food environment crisis may suffer physical distress, leading them to rely on opioid-containing painkillers and, ultimately, succumb to addiction and death. Lena is determined to aid this vulnerable population, utilizing CausalChat's Create Sub-Graph module to isolate these variables and associated edges, creating a new child node in the model tree. Her objective now is to identify efficient direct factors that can influence this causal pathway. Using Latent Factor Charts (not shown), Lena identifies three intervenable measures to improve this pathway: economic incentives for healthy food retailers, physical activity promotion, and prescription drug monitoring programs (see Fig. 16). However, despite the rationale behind these latent factors, their effectiveness in reducing opioid death rates cannot be conclusively determined and would require research with real data and experimentation. At this juncture, they are merely conceptualized ideas facilitated by Lena's utilization of CausalChat.

V. User Study

We conducted a two-fold user study: (1) Obtaining feedback from domain experts to assess the practical applicability and logical coherence of our proposed framework; (2) Evaluating the usability and efficacy of our framework by assigning tasks to non-expert users.

A. Datasets

In these user studies, we utilized two real-world datasets related to public health. Each data point corresponds to a distinct county in the United States for the year 2019. The expert assessment utilized the opioid death dataset described in section IV. The non-expert user study employed the Life Expectancy dataset detailed below.

The Life Expectancy dataset comprises 8 key variables sourced from the County Health Rankings & Roadmaps Database [45]: firearm fatality rate, violent crime rate, average grade performance, high school graduation rate, food environment index, percent fair or poor health, primary care physician rate and debt income ratio for each of more than 3,000 US counties, All of these variables are recognized to affect demographic life expectancy either directly or indirectly.

B. Expert Assessment

We conducted a qualitative expert study to assess the value of CausalChat for scientific work. Through semi-structured interviews, we explored how scientists used the tool in familiar domains under two usage modes, distinguished by their trust in automated causal inference. We aimed to gather feedback on the tool's potential value for their research and identify actionable suggestions for future development. To that end, we invited three domain experts from our university—specializing in health policy (P1), exposure science (P2), and environmental epidemiology (P3)—all of whom have strong backgrounds in causal inference and a shared interest in applying quantitative methods to public health modeling.

Each of the three sessions took place via Zoom. We first introduced our interface, emphasizing the functions and significance of each visual component. We ensured that the experts were familiar with how to navigate the system. Subsequently, they were tasked with refining a causal graph containing two unresolved edges and one misdirected edge, with the goal to achieve a valid causal graph, and eventually ideating additional factors. All were able to achieve these tasks. Throughout their interaction with the system, we encouraged them to articulate their thought process. Finally, we gathered their feedback on CausalChat, specifically soliciting suggestions for potential enhancements. In the following we grouped the verbal session outcomes into specific themes.

Overall assessment. All three participants unanimously praised CausalChat as an exceptional ideation tool, noting its effectiveness in helping non-experts quickly grasp the most prominent causal relations between variables in a field potentially unfamiliar to them. Specifically, P2 highlighted the efficiency of the tool, noting, "CausalChat enlightens non-expert users in a productive fashion. Users don't have to go through an exhaustive literature research process to obtain a fundamental understanding of a new field."

Validation. P1 expressed that her expertise led her to anticipate the presence of *Withdrawal Treatment* as a mediator between opioid dispensing and opioid-related deaths. This anticipation was validated when she discovered *Withdrawal Treatment* listed as a significant mediator within the Causal Relationship Environment chart. She added that once identified conceptually, the proposed mediators, confounders, and other new variables can be statistically tested, enhancing confidence and adding accuracy beyond GPT-4's strength assessments.

Making access to domain science easier and more streamlined. The potential of our tool to provide expedited access to domain knowledge became evident when the experts confirmed CausalChat's adeptness in identifying confounders and mediators within an existing causal graph. P1 remarked that "compared to the traditional method of studying and including all possible confounders, which demands relentless and tedious literature review, CausalChat offers a far more efficient solution by automatically and visually presenting all potential confounders. This feature significantly enhances accessibility to science".

Use as a research tool. P3 perceived CausalChat as an actual research planner. She said: "I would use the framework for planning my research. With all these essential variables and their information visually represented, I can have a clear goal of what data I need to collect, what research papers are useful, and what academic areas I can potentially contribute to." She further noted that the synergy between the Causal Debate Chart and the Causal Relation Environment Chart not only highlighted well-studied areas but also suggested intervention strategies (such as mediators) and ways to mitigate unrealistic relationships. Relatedly, P1 remarked that the framework effectively prompts domain experts to consider essential components that may be overlooked but are crucial for enhancing the outcome variable.

Preferred alternative to pure data-driven causal analysis. P2, an expert in experimental-based causality, began his session by voicing strong skepticism with regards to existing methods for automated causal analysis. He contended that observational data could only uncover correlations, not causation. According to his view external information is indispensable for directing causal edges, and even for forming them in the first place. He favors the alternative approach supported by CausalChat – taken by Lena in our second case study – which first constructs and refines a causal graph by leveraging the knowledge of GPT-4 and then estimates the causal effect of each edge using data.

Satisfying the need for personalized causal models. P2 emphasized that socioeconomic variables are predominantly cross-sectional and can engender feedback loops, resulting in bidirectional edges. This poses a challenge to the assumption of the causal graph being a DAG, underscoring the importance of further expanding our model tree panel. In each distinct causal model, there exists a distinct data generating process where variables have a clear upstream and downstream direction. An edge pointing from A to B in one model can point reversely in another. These two models will share a parent model in our model tree. After learning about this feature, P2 acknowledged its effectiveness in addressing his concern.

Concerns with GPT-4's quality of citations. P1 and P3 expressed concerns about the inherent uncertainties of GPT-4. CausalChat aims to provide citations whenever clarifying a standpoint or evaluating a causal relation. However, when relevant sources are unavailable, typically because the causal hypothesis has not been extensively studied, the evaluation

about the causal relation depends on the GPT-4's inference. Additionally, GPT-4 may struggle to distinguish rigorous from non-rigorous research due to varying literature quality. While this may not pose a problem for well-trained scientists who can filter out moderate papers, there is concern that users might rely on CausalChat for decision-making and consider it the ultimate truth. P1 suggested including a disclaimer indicating that some edges may be misdirected or omitted.

C. Non-Expert User Study

We also conducted a quantitative study to evaluate CausalChat's usability and effectiveness with users lacking domain expertise (see Appendix A for a complete breakdown of the study's results). We aimed to show that CausalChat's benefits stem from its design, not merely from the use of LLMs. To this end, we implemented an ablation study that progressively added key components: statistical feedback (BIC), LLM-generated text, and the full interactive visual interface. Each baseline isolates the contribution of a specific component, allowing us to assess how the integration of LLMs and visual reasoning tools improves causal exploration. Participants were evaluated on their ability to understand interrelationships among variables, complete causal auditing tasks efficiently, and rate the system's ease of use—measured through a combination of task performance and questionnaire scoring. Our study tested the following three hypotheses:

- **H1.** CausalChat provides comprehensive yet concise guidance to help users uncover interrelationships among variables. It also encourages users to identify latent third variables, such as confounders and mediators, that may influence the relationship between two variables.
- **H2.** CausalChat improves users' efficiency in conducting causal analysis in unfamiliar domains, particularly by supporting the identification of interrelationships and latent variables, as in H1.
- **H3.** CausalChat is user-friendly, convenient, and effective.

To test these hypotheses, we conducted an ablation study with three layers. The first layer relies on traditional causal model reasoning using the BIC score as a quality metric, the second layer incorporates text prompts for an LLM, and the third layer integrates these methods into the CausalChat interface. Each layer progressively adds more sophisticated tools to enhance causal reasoning and decision-making.

- **L1. Conventional BIC-Score Based Causal Reasoning:** In this layer, users refine causal graphs by combining their domain knowledge with feedback on modifications, evaluated through the BIC score.
- **L2. LLM-support with text:** This layer builds on L1 by incorporating a ChatGPT interface. After specifying the relationship of interest, users receive exemplar prompts for text responses to explore causal connections, allowing them to determine the most plausible causal relation based on multiple GPT-4 responses addressing various aspects of the relationship.
- **L3. CausalChat Lite:** This layer adds to L2 some of the visual elements of CausalChat: (1) the causal debate charts, (2) the causal relation environment charts, and (3) the justification panel.

Our study focuses on CausalChat's ability to correct distorted effects, identify mediation effects for indirect relationships, and resolve edge directionality in causal graphs. The aim is not to encourage users to endlessly expand the graph but to guide them in accurately determining the direction of undirected edges and addressing omitted variables that are critical within the scope of the theme under consideration.

Dataset. To minimize variation in the final graph, we used a subset of the Life Expectancy dataset. This subset includes the following variables: food environment index, percent in fair or poor health, primary care physician rate, debt-to-income ratio, and life expectancy.

Participants. We recruited six university students (3 males, 3 females) for usability studies. All participants were familiar with web browser-based frameworks and could participate either in person or via Zoom. None of the participants had expertise in epidemiology.

Study Design. Participants were guided through the study using Qualtrics, an online survey tool that follows a predefined workflow. The workflow began with tutorials on data causality and instructions for using CausalChat, followed by a quiz to reinforce participants' understanding of graphical causal models, including terms related to confounders and mediators. Participants had to answer all quiz questions correctly before proceeding to the main tasks involving CausalChat evaluation. Finally, they completed a series of usability questions, including the standardized System Usability Scale (SUS) questionnaire, to provide feedback on their experience.

Ablation Study Stages. We divided a session into three stages, each corresponding to the testing of a different layer (method). At each stage, participants were asked to resolve undirected edges using the tools available for that stage. For each undirected edge, participants had the option to direct the edge, remove it, add a confounder, or add a mediator. Initially, participants were presented with 9 preset undirected edges derived from the subset of variables in the life expectancy dataset. To ensure fairness, participants could resolve any 3 undirected edges of their choice at each stage until they were satisfied. They were not allowed to modify any edges they had already edited in previous stages.

Stage 1. In the first stage, participants were asked to resolve 3 undirected edges using BIC score feedback. A bar displayed the change in the BIC score for the affected nodes after each edge modification, enabling participants to assess the impact of their changes.

Stage 2. In the second stage, participants were asked to resolve another 3 undirected edges of their choice using LLM support provided through a set of pre-formulated prompts, similar to those used by CausalChat and consisting of debate prompts and causal relation environment prompts. Participants received text-based responses from GPT-4 based on these prompts and made their decisions by reviewing the responses.

Stage 3. In the final stage, all operations were conducted within CausalChat Lite. For the last 3 edges, participants made decisions based on the visual charts and verified their insights by reviewing the corresponding justifications.

H1: Understanding Interrelationships Among Variables

Stage 1. In the first stage, participants based their decisions largely on their own knowledge and the BIC score. However, without domain expertise, their decisions were often uncertain or biased. Although the BIC score provided insights into key predictors of the target variable, it was vulnerable to distortions from hidden confounders, which led to inaccuracies in identifying true causal effects. As a result, 28% (5/18) of the undirected edges were incorrectly resolved. Participants also identified only one mediator and no confounders. When asked why they struggled with latent variables, participants pointed to two main challenges: they tended to focus on relationships they felt most confident about, which were often direct or unrelated, and their lack of expertise made it difficult to recognize potential confounders and mediators, limiting their ability to discover these variables effectively.

Stage 2. In the second stage, introducing GPT-4 as a proxy for domain expertise helped participants address knowledge gaps from stage one. However, participants' performance declined slightly, with 39% (7/18) of causal edges resolved incorrectly, compared to 5/18 in the BIC score method. This is likely because they tackled the easier edges first in stage one. Nevertheless, we observed that the volume of GPT-4 responses proved somewhat overwhelming, making it difficult for participants to fully comprehend and synthesize the information, despite GPT-4's solid understanding of causal terminology, minimal hallucinations, and useful references to relevant literature. We observed that this guidance was particularly helpful in determining causal directionality and encouraging a deeper exploration of potential confounders and mediators. Although participants struggled at times to process the information, they identified four valid confounders and six valid mediators, with none deemed incorrect. While latent variables remained difficult to identify, the overall decision quality showed a clear improvement compared to stage one.

Stage 3. CausalChat Lite performed the best overall, with only 1/18 incorrect edges (5%), even though these were the most difficult relations since this stage came last. Participants found it significantly easier to navigate the tasks. For example, P2 noted: "It becomes much easier after realizing how straightforward it is when the patterns are visualized in the causal debate chart. Before, it was really hard to memorize all the queries at once." The chart simplified the process by visually representing complex relationships, reducing the cognitive load of managing multiple queries simultaneously.

We observed that participants made good use of the debate justification text box to obtain clear reasoning for each causal hypothesis rating. The causal relation environment chart further supported participants by visualizing confounders and mediators together, offering quick access to potential third variables relevant to their analysis. Across the study, participants identified seven valid confounders and five valid mediators, with no false positives, underscoring the effectiveness of CausalChat in improving the accuracy of causal analysis.

H2: Efficiency in Causal Auditing Tasks

We assessed the efficiency of CausalChat by measuring two key aspects of the causal auditing tasks: (1) edge modification and (2) the discovery of confounders and mediators. As discussed in the analysis of **H1**, BIC score based analysis alone is insufficient for guiding users to construct accurate causal graphs. Therefore, our focus here is on comparing the causal auditing efficiency between the LLM-assisted interactions and CausalChat Lite.

For edge editing, we measured the time from when a participant began querying the relationship between two variables until they verbally confirmed satisfaction with the modification. With the LLM support, the length and complexity of responses made reading time-consuming and, for some, tedious. As a result, participants spent an average of 3.3 minutes (SD = 2.6) using the LLM text, while the time decreased to 0.9 minutes on average (SD = 0.9) with CausalChat Lite.

Discovering confounders and mediators involved three steps: (1) querying potential confounders and mediators for a variable pair, (2) evaluating the logical consistency of the identified variables, and (3) adding valid variables to the causal graph. On average, for the LLM-supported version, participants spent 3.5 minutes (SD = 0.5) to add a confounder and 3.6 minutes (SD = 2.1) to add a mediator. Participants spent significantly less time for this task with CausalChat Lite. On average, it took 1.6 minutes (SD = 1.0) to add a confounder and 1.3 minutes (SD = 0.4) to add a mediator. This suggests that CausalChat not only improves efficiency in edge editing but also streamlines the process of identifying and incorporating latent variables.

H3: Easy to Use

Participants unanimously agreed that CausalChat Lite was the most convenient (M = 4.8, SD = 0.4) and instilled the most confidence (M = 4.2, SD = 0.7) compared to the BIC Score based analysis (BIC) and the LLM text support (LLM). The average confidence score for BIC was 2.5 (SD = 0.8), and 3.3 (SD = 0.7) for LLM. In terms of convenience, BIC scored 2.7 (SD = 1.1), while LLM scored 3.3 (SD = 1.2). CausalChat Lite also achieved a usability score of 79.17 (SD = 6.40) on the standardized SUS questionnaire, placing it in the 85th to 89th percentile—well above the average SUS score of 68 and close to the top 10% of all SUS scores (80.8).

These results aligned with our expectations for CausalChat in assisting users with causal relation auditing, causal graph refinement, and decision-making. In post-study interviews, participants praised CausalChat as a comprehensive tool for quickly learning about unfamiliar fields. P2 remarked, 'CausalChat did an efficient job condensing the wordiness of GPT-4 responses into more digestible visual infographics, making decision-making easier.' Participants also appreciated the visual designs of the causal debate chart and the causal relation environment chart, noting that once understood, learning causal relationships became much smoother. P4 stated, 'The visualizations are very helpful, especially with the color coding to reflect the direction of mediators or confounding variables.'

VI. Discussion

While LLMs can produce insightful causal reasoning, they also exhibit well-known limitations, such as hallucinations and difficulty handling ambiguity. CausalChat addresses these

challenges by integrating a robust suite of textual and visual features that make the model's reasoning transparent and interactive. These features are essential for bridging the gap between AI-generated suggestions and human understanding, allowing users to probe, validate, and refine causal claims. Our case studies demonstrate CausalChat's potential to enhance user-driven innovation, support the development of more accurate causal models, and advance the study of complex systems across a range of domains.

Yet several limitations remain. First, there is the current inability to directly represent feedback loops or causal relationships whose direction depends on initial conditions—such as the relationship between sleep deprivation and stress. Some causes are also better modeled as moderators or contextual conditions; for example, trees are not the cause of forest fires per se, but a necessary precondition. Both of these issues could potentially be addressed through a more sophisticated version of our causal model tree, which is still in its early stages. Second, ChatGPT currently lacks the ability to provide reliable citations for its justifications—though this limitation appears to be improving as newer models come online.

An interesting direction for future work is introducing temporal or personalized context into prompts. For example, Oscar, the automotive hobbyist, could explore causal networks across time—rolling back to the 1920s or imagining a future dominated by electric vehicles—by priming the LLM with historical or speculative context. Similarly, personalization could allow Lena, the epidemiologist, to prompt the LLM for a specific population group and so generate causal models tailored to these specific populations, such as rural blue-collar workers, enabling targeted health interventions.

VII. Conclusions

We presented CausalChat, a system that leverages the causal knowledge of large language models (LLMs), particularly GPT-4, to make advanced causal insights accessible to those innovating on complex systems. Our approach addresses skepticism about automated causal inference, which often stems from the challenge of acquiring datasets comprehensive enough to model such systems. Instead of relying on large datasets, we directly query LLMs using carefully crafted prompts. User studies, including those involving participants skeptical of automated causal inference, showed that our LLM-powered tool effectively meets their needs.

Future work will focus on enhancing the scalability of CausalChat's visual interface to support large causal graphs representing complex models, most likely via a level-of-detail approach. In addition, having an insight log associated with each variable and link would allow users to preserve insight gained from the justification panel. Further, we aim to enhance support for periodic causal model validation—not only through SEM for parameter estimation, but also by checking the structure itself using GES and appropriate statistical tests to assess model robustness and consistency with available data, for example through Bayesian Additive Regression Trees (BART) [47]. Additionally, we aim to develop a crawling mechanism capable of automatically retrieving data for newly added variables and causal relations from online repositories.

Acknowledgments

This research was funded in part by CDC Cooperative Agreement 1 NH25PS005202-01-00, American Public Health Association (APHA) award # 2023-0004, and NSF grant CNS 1900706.

References

[1] D. Hume, *An Enquiry Concerning Human Understanding and Other Writings*. Cambridge University Press, 2007.
[2] I. Kant, "Critique of pure reason. 1781," *Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin*, pp. 370–456, 1908.
[3] J. Pearl, M. Glymour, and N. Jewell, *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.
[4] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.
[5] C.-H. Yen, H. Cheng, Y. Xia, and Y. Huang, "Crowdidea: Blending crowd intelligence and data analytics to empower causal reasoning," in *ACM CHI*, 2023, pp. 1–17.
[6] E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," *arXiv preprint arXiv:2305.00050*, 2023.
[7] J. Vandenbroucke, A. Broadbent, and N. Pearce, "Causality and causal inference in epidemiology: the need for a pluralistic approach," *Intern. Journal of Epidemiology*, vol. 45, no. 6, pp. 1776–1786, 2016.
[8] T. Glass, S. Goodman, M. Hernán, and J. Samet, "Causal inference in public health," *Annual Review of Public Health*, vol. 34, pp. 61–75, 2013.
[9] T. Dang, P. Murray, J. Aurisano, and A. Forbes, "Reactionflow: An interactive visualization tool for causality analysis in biological pathways," in *BMC Proceedings*, vol. 9, 2015, pp. 1–18.
[10] J. Gerring, "Causation: A unified framework for the social sciences," *Journal of Theoretical Politics*, vol. 17, no. 2, pp. 163–198, 2005.
[11] J. Wang and K. Mueller, "The visual causality analyst: An interactive interface for causal reasoning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 230–239, 2015.
[12] ——, "Visual causality analysis made practical," in *IEEE VAST*, 2017, pp. 151–161.
[13] X. Xie, F. Du, and Y. Wu, "A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1448–1458, 2020.
[14] P. Spirtes, C. Glymour, R. Scheines, and D. Heckerman, *Causation, Prediction, and Search, Second Edition*. MIT Press, 2000.
[15] P. Spirtes, "An anytime algorithm for causal inference," in *International Workshop on Artificial Intelligence and Statistics*, 2001, pp. 278–285.
[16] D. Chickering, "Optimal structure identification with greedy search," *Journal of Machine Learning Research*, vol. 3, pp. 507–554, 2002.
[17] K. Burnham and D. Anderson, "Multimodel inference: understanding aic and bic in model selection," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, 2004.
[18] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, pp. 461–464, 1978.
[19] R. Kline, *Principles and practice of structural equation modeling*. Guilford publications, 2023.
[20] C. Shalizi, *Limitations on Consistency of Causal Discovery*. New York, NY, USA: Cambridge University Press, 2013, ch. 22.6.
[21] J. Robins, R. Scheines, P. Spirtes, and L. Wasserman, "Uniform consistency in causal inference," *Biometrika*, vol. 90, no. 3, pp. 491–515, 2003.
[22] J. Yang, S. Han, and J. Poon, "A survey on extraction of causal relations from natural language text," *Knowledge and Information Systems*, vol. 64, no. 5, pp. 1161–1186, 2022.
[23] S. Long *et al.*, "Can large language models build causal graphs?" *arXiv preprint arXiv:2303.05279*, 2023.
[24] E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," *arXiv preprint arXiv:2305.00050*, 2023.
[25] A. Nam, C. Hughes, T. Icard, and T. Gerstenberg, "Show and tell: Learning causal structures from observations and explanations," in *Proc. Annual Conference of the Cognitive Science Society*, 2023.
[26] S. Long, A. Piché, V. Zantedeschi, T. Schuster, and A. Drouin, "Causal discovery with language models as imperfect experts," *arXiv preprint arXiv:2307.02390*, 2023.

[27] J. Gao, X. Ding, B. Qin, and T. Liu, "Is chatgpt a good causal reasoner? a comprehensive evaluation," *arXiv preprint arXiv:2305.07375*, 2023.
[28] Z. Jin *et al.*, "Can large language models infer causation from correlation?" *arXiv preprint arXiv:2306.05836*, 2023.
[29] M. Zečević, M. Willig, D. Dhami, and K. Kersting, "Causal parrots: Large language models may talk causality but are not causal," *arXiv preprint arXiv:2308.13067*, 2023.
[30] D. Berenberg and J. Bagrow, "Efficient crowd exploration of large networks: The case of causal attribution," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–25, 2018.
[31] S. Salim, N. Hoque, and K. Mueller, "Belief miner: A methodology for discovering causal beliefs and causal illusions from general populations," *Proc. ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–37, 2024.
[32] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *ACM CHI Extended Abstracts*, 2021, pp. 1–7.
[33] C.-H. Yen *et al.*, "Narratives+ diagrams: An integrated approach for externalizing and sharing people's causal beliefs," *Proc. ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–27, 2021.
[34] J. Textor, J. Hardt, and S. Knüppel, "Dagitty: a graphical tool for analyzing causal diagrams," *Epidemiology*, vol. 22, no. 5, p. 745, 2011.
[35] J. Wang and K. Mueller, "Domino: Visual causal reasoning with time-dependent phenomena," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 12, pp. 5342–5356, 2023.
[36] G. Guo, E. Karavani, A. Endert, and B. Kwon, "Causalvis: Visualizations for causal inference," in *ACM CHI*, 2023, pp. 1–20.
[37] Y. Zhang *et al.*, "An explainable ai approach to large language model assisted causal model auditing and development," in *Proc. NL-VIZ Workshop at IEEE VIS 2023*, Melbourne, Australia, 2023.
[38] X. Wang *et al.*, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.
[39] S. Mishra, D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi, "Reframing instructional prompts to gptk's language," *arXiv preprint arXiv:2109.07830*, 2021.
[40] M. Sharma, H. C. Siu, R. Paleja, and J. D. Peña, "Why would you suggest that? human trust in language model responses," *arXiv preprint arXiv:2406.02018*, 2024.
[41] P. Tennant *et al.*, "Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations," *Intern. Journal of Epidemiology*, vol. 50, no. 2, pp. 620–632, 2021.
[42] E. Murray and Z. Kunicki, "As the wheel turns: Causal inference for feedback loops and bidirectional effects," OSFPreprints, 7 2022.
[43] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
[44] R. Quinlan, "Auto MPG," UCI Machine Learning Repository, 1993, DOI: https://doi.org/10.24432/C5859H.
[45] University of Wisconsin Population Health Institute, "County health rankings & roadmaps," www.countyhealthrankings.org, 2023.
[46] National Center for Health Statistics, "Cdc wonder online database," http://wonder.cdc.gov/ucd-icd10-expanded.html, 2021.
[47] J. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

Yanming Zhang is currently a PhD student in Computer Science at Stony Brook University. He obtained a Bachelor degree in Mathematics and Applied Mathematics from Southwest Jiaotong University. His research focuses on explainable AI, visual analytics, causal inference, and human-computer interaction. For more information, see https://yanmluk.github.io

Akshith Reddy Kota holds a MS in Computer Science, Stony Brook University. His research interests include machine learning, deep learning, explainable AI, visual analytics, and big data analytics. He earned a Bachelor of Technology in CSE from Vellore Institute of Technology, India.

Eric Papenhausen holds a PhD in Computer Science, Stony Brook University. His research interests includes machine learning, deep learning, explainable AI, visual analytics, big data analytics, and medical image synthesis. He is currently CTO at Akai Kaeru LLC.

Klaus Mueller is currently a professor of Computer Science at Stony Brook University and a senior scientist at Brookhaven National Lab. His research interests include explainable AI, visual analytics, data science, and medical imaging. He won the US NSF Early CAREER Award and has co-authored $>$ 300 papers, cited more than 14,500 times. He is a Fellow of the IEEE.