

From Language to Causality: Extracting Causal Relations from Large Language Models

Márk Marosi, Kristóf Váradi, Péter Antal
Budapest University of Technology and Economics
Department of Artificial Intelligence and Systems Engineering
Budapest, Hungary
Email: {marosi, antal}@mit.bme.hu, kristofvaradi@edu.bme.hu

Abstract—This research introduces a novel framework for constructing causal networks by leveraging the causal reasoning abilities of multiple Large Language Models (LLMs). We instruct LLMs to extract explicit causal links from their internal knowledge representations regarding specific topics. We explore methods for consolidating these graphs, addressing conflicts, and determining the strength and directionality of causal links. Evaluated across various domains using the Qwen 2.5 model family (0.5B to 14B parameters), the framework demonstrates the ability of language models to generate meaningful causal networks from complex queries. Our findings suggest that fusing causal knowledge from multiple LLMs significantly enhances causal discovery from natural language, though practical application benefits from human oversight and domain expertise to ensure accuracy and reliability. We also highlight the potential of integrating probabilistic approaches to quantify uncertainty within the extracted causal relationships.

Index Terms—Large Language Models, Natural Language Processing, Bayesian Networks, Causal Discovery, Probabilistic Graphical Models

I. INTRODUCTION

Causal Bayesian networks, represented as directed acyclic graphs (DAGs), have become indispensable tools for modeling causal relationships across diverse scientific disciplines, including medicine, engineering, and social sciences [12]. Traditionally, constructing these models relies on expert elicitation, statistical analysis of observational data, or a combination thereof [4]. However, the emergence of Large Language Models (LLMs) offers a compelling new paradigm: extracting causal knowledge directly from their vast latent representations. This approach shares common goals with foundational work on automated knowledge discovery from scientific literature [13, 14], and extends text-mining and natural language processing methods that integrate textual information with other data modalities for knowledge synthesis [9, 5, 8, 1, 2]. While our approach similarly aims for extracting causal relations from scientific literature used in LLMs training, it differs fundamentally in that we are not limited to relations that are explicitly mentioned in the text but leverage the LLM’s ability to generate plausible causal relations not seen in the data. This hypothesized capacity of

LLMs’ latent representations correspond to the causal level of their semantic compositionality.

We introduce a framework that leverages multiple LLMs to infer causal relations and complete causal diagrams from their latent representations, as expressed through their textual outputs. Our methodology, visualized in Figure 1, comprises three key stages. First, we prompt multiple LLMs to generate causal subgraphs related to a specific topic, encouraging diverse perspectives and mitigating individual model biases. We guide the generation process toward tree-like structures initially to reduce the risk of cyclical dependencies. Second, we employ various post-processing techniques to refine and unify the extracted graphs. We develop strategies to reconcile conflicting causal assertions, leveraging the collective wisdom of multiple LLMs. Semantic similarity measures, based on embedding vectors, are used to merge semantically equivalent nodes, thereby reducing redundancy. We also implement graph consolidation procedures that further refine the network representation. Finally, we investigate methods for assessing the existential certainty and directionality of causal relationships using the aggregated knowledge of the LLMs.

Recent research has begun to explore the potential of LLMs for causal reasoning tasks [6, 16, 15, 10, 7, 3]. Our work distinguishes itself by focusing on the explicit construction and refinement of causal network structures by synthesizing knowledge from multiple LLMs. We present a comprehensive evaluation across multiple domains using the Qwen 2.5 model family (0.5B to 14B parameters) [17]. Our results confirm the emergence of causal compositionality in LLM latent representations and demonstrate the utility of this approach for automated causal knowledge discovery and highlight key challenges and future research directions at the intersection of LLMs, causal discovery, and causal inference. We emphasize the importance of human-in-the-loop validation and the advantage of incorporating probabilistic reasoning to manage uncertainty in the extracted causal relations.

Framework for Causal Network Construction from LLM Knowledge

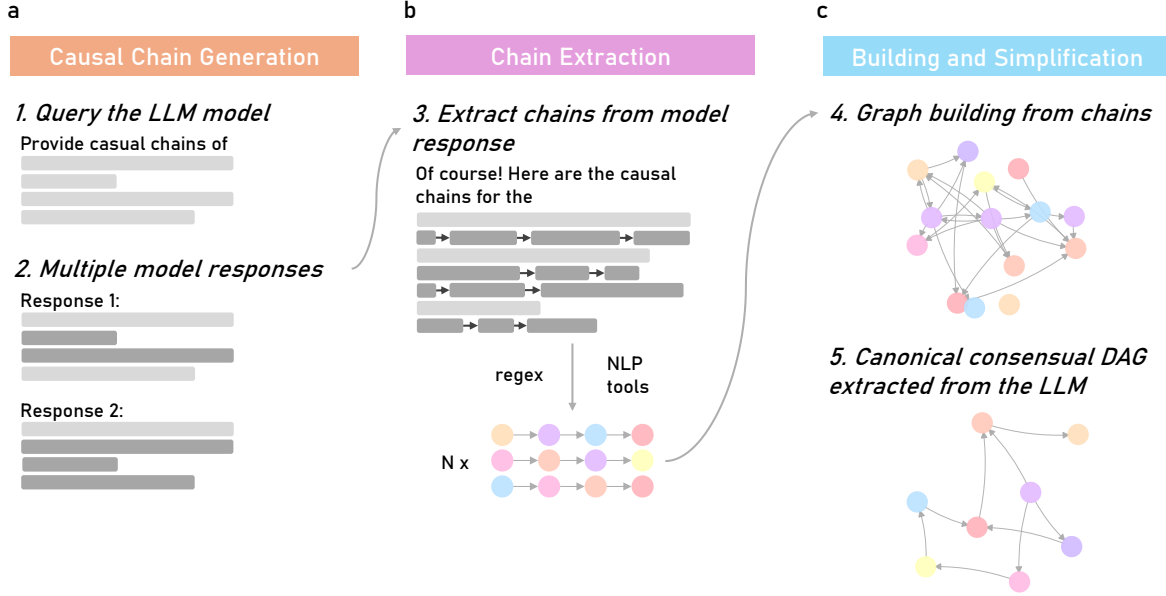


Fig. 1. Framework for constructing causal networks from large language models. The process comprises three stages: (a) **Causal Chain Generation:** The LLM is queried with a prompt to elicit multiple responses detailing causal relationships. (b) **Chain Extraction:** Natural language processing (NLP) tools and regular expressions (regex) are employed to systematically extract structured causal chains from the LLM’s textual outputs. (c) **Building and Simplification:** The extracted chains are used to construct an initial causal graph. This graph is then simplified to obtain a canonical directed acyclic graph (DAG), representing the consensus causal structure derived from the LLM’s knowledge. The framework leverages natural language understanding, graph-based representation, and consensus-building methods to generate a robust causal network.

II. BACKGROUND

A Bayesian Network (BN) is a probabilistic graphical model that represents a set of random variables $\{X_1, X_2, \dots, X_n\}$ and their conditional dependencies via a directed acyclic graph. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed acyclic graph (DAG), where: $\mathcal{V} = \{X_1, X_2, \dots, X_n\}$ is the set of nodes, each representing a random variable and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of directed edges, indicating direct dependencies between variables.

The joint probability distribution of the variables $\{X_1, X_2, \dots, X_n\}$ in the network factorizes according to the structure of the graph \mathcal{G} :

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \text{Pa}(X_i)),$$

where $\text{Pa}(X_i)$ denotes the set of parent nodes of X_i in the graph \mathcal{G} . Each node X_i is associated with a conditional probability distribution (CPD) $P(X_i \mid \text{Pa}(X_i))$, which specifies the probabilities of X_i given its parent nodes.

Bayesian Networks satisfy the Parental Markov property, which states that a node X_i is conditionally independent of its non-descendants given its parent nodes:

$$X_i \perp\!\!\!\perp \text{NonDescendants}(X_i) \mid \text{Pa}(X_i).$$

Although Bayesian Networks are often used to model causal relationships, a directed edge from u to v does not necessarily imply that X_v is causally dependent on X_u [11]. A causal network is a Bayesian Network that explicitly encodes causal relationships (mechanisms) [12]. In a causal network, performing an intervention, denoted by $\text{do}(X = x)$, modifies the network by removing the edges into X and setting X to the value x . This allows for predicting the effects of external interventions using the induced truncated probability distributions.

III. METHODS

A. Network Extraction

To address the challenge of cycle creation and incoherence in LLM-generated Bayesian Networks, we introduce a network extraction approach using directed subgraphs. We

prompt LLMs to produce causal “chains” (e.g., $A \rightarrow B \rightarrow C$) with few-shot examples guiding structured output like `Cause -> Effect -> ...`. While we prompt the LLMs to generate chains in a structured $A \rightarrow B \rightarrow C$ format, variations in responses can occur. To ensure robust extraction, we employ a regular expression, specifically `'([\w\s.-]+)\s*->\s*([\w\s.-]+)'`, designed to capture cause-effect pairs even if minor deviations from the prompted format are present. This regex accommodates variable names containing spaces and special characters. This regular expression is crucial because, despite few-shot prompting, LLMs may not always adhere strictly to the desired output structure. The regex provides a reliable mechanism to extract the intended causal relationships from the LLM’s free-form text. These pairs are then merged into an unweighted causal graph, where branching chains (e.g., $A \rightarrow B$, $A \rightarrow C$) are unified under a single root.

B. Existential Uncertainty of Causal Relations

To quantify the degree to which an LLM supports a given causal relation $A \rightarrow B$, we introduce a **certainty score** derived from the model’s token-level output probabilities. This score reflects the model’s confidence in B being a consequence of A . We treat the effect, B , as a sequence of tokens because we are measuring the model’s confidence in generating B given A . By summing the log probabilities of each token in B , we estimate the overall likelihood of the model producing B in response to A . The cause, A , is part of the input prompt and is not treated as a sequence of tokens for this calculation.

Formally, given a causal relationship $A \rightarrow B$, we construct a context string:

$$\text{context} = \text{concatenate}(\text{prompt}, A, " \rightarrow ")$$

and then measure the log probability of generating the sequence of tokens in $B = (b_1, b_2, \dots, b_m)$. The certainty score for the directed edge $A \rightarrow B$ is:

$$\text{Certainty}(A \rightarrow B) = \sum_{i=1}^m \log p(b_i \mid \text{context}, b_1, \dots, b_{i-1}).$$

Here, $p(b_i \mid \dots)$ is the conditional probability the LLM assigns to token b_i given the prefix. The certainty score for a given edge $A \rightarrow B$ is defined as the sum of the log probabilities of the tokens in B , given the context (prompt + A + " \rightarrow ") and the preceding tokens of B .

In cases where multiple extractions from different chains or models yield the same pair (A, B) , we retain the maximum certainty score among duplicates. This reflects the strongest association between A and B as a likely cause-effect pair according to the LLM’s learned representations.

C. Conflict Resolution for Opposite Edges

Occasionally, our pipeline uncovers contradictory statements such as $A \rightarrow B$ and $B \rightarrow A$. To resolve these conflicts, we aggregate the certainty scores for each direction. If one direction has a substantially higher score, we select that direction. If the scores are comparable, we label the edge as *ambiguous* and flag it for human review. When merging nodes results in conflicting edges (e.g., $A \rightarrow B'$ and $D \rightarrow B'$ after merging B and C), we aggregate the certainty scores as follows. For each direction, we sum the certainty scores of the original edges that contributed to the merged edge.

In the example [1] $A \rightarrow B \rightarrow D \rightarrow E$; [2] $A \rightarrow C \leftarrow D \rightarrow E$; $B + C = B'$, we would calculate:

- $\text{Certainty}(A \rightarrow B') = \text{Certainty}(A \rightarrow B)$
- $\text{Certainty}(B' \leftarrow D) = \text{Certainty}(C \leftarrow D)$

We then compare these aggregated scores to determine the final edge direction, following the same procedure as for non-merged conflicting edges. If the certainty scores are not significantly different, we mark the edge $A \leftrightarrow B$ as *ambiguous*.

D. Post-processing

After extracting cause-effect pairs, many nodes represent the same concept but use different wording. We reduce redundancy by merging nodes whose cosine similarity exceeds a threshold τ . We obtain embedding vectors for node labels using the pre-trained Sentence-BERT model, specifically the `'all-mpnet-base-v2'` model, which has demonstrated strong performance in semantic similarity tasks. Throughout our experiments, we set $\tau = 0.8$.

We find that increasing τ above 0.9 under-merges nodes (e.g., “global warming” and “climate change” might remain separate), while lowering τ below 0.7 tends to over-merge semantically related but distinct concepts (e.g., “greenhouse gases” might be merged with “climate change”). We set the cosine similarity threshold τ to 0.8 based on empirical evaluation and common practice in textual similarity tasks. The method’s sensitivity to τ is moderate; small variations around 0.8 do not drastically alter the results, but larger deviations can lead to significant under- or over-merging, impacting the final network structure.

IV. EXPERIMENTS

A. Network Complexity

To evaluate LLMs’ ability to identify causal relationships across diverse subjects, we develop a suite of 50 few-shot tasks spanning multiple domains with various topics and corresponding target variables. Our evaluation considers 5 advanced topics in each of *Healthcare*, *Environmental Science*, *Computer Science*, *Physics*, *Biology*, *Chemistry*, *Mathematics*, *Psychology*, *Engineering*, and *Astronomy*. In this work, we focus on evaluating the size and complexity

Scaling of Num Edges with Model Size by Domain

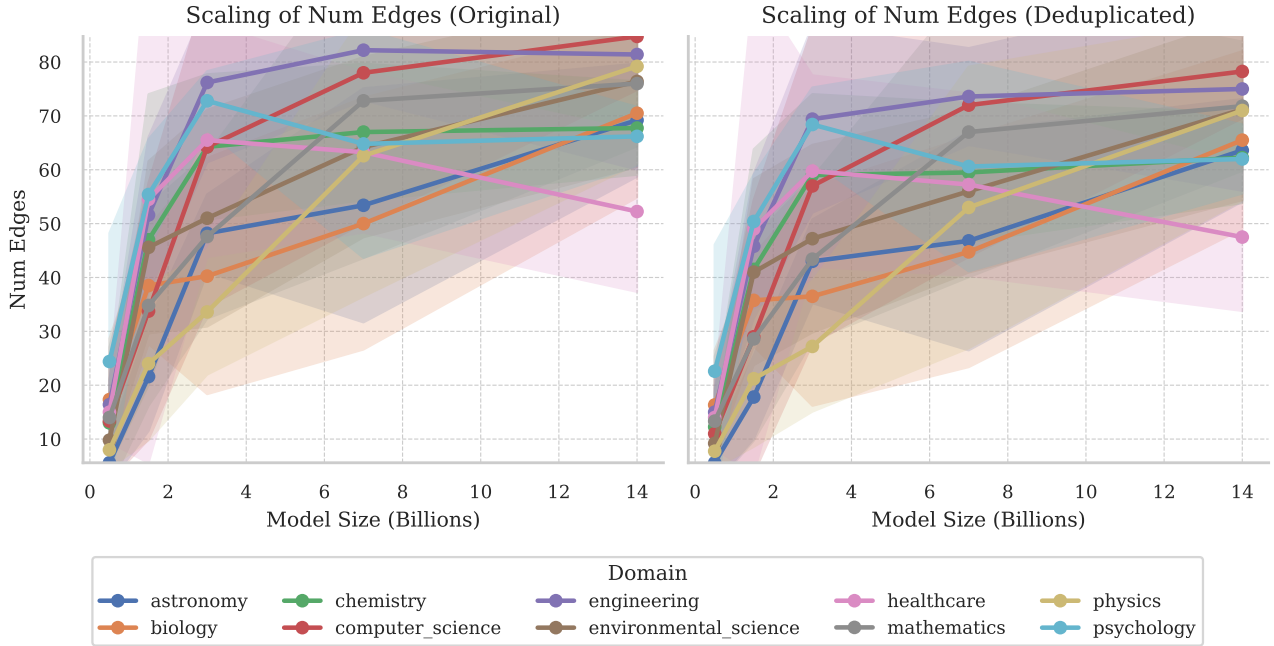


Fig. 2. Scaling of the number of edges in causal graphs with increasing model size across different domains. The figure presents two line plots illustrating the relationship between the size of the LLM used (in billions of parameters) and the number of edges in the resulting causal graphs. The left plot shows the scaling for the "original" graphs, while the right plot depicts the scaling for the "deduplicated" graphs. The shaded area around each line represents the variance across topics runs. A general trend of increasing edge count with larger model sizes is observed across most domains in both the original and deduplicated graphs, suggesting that larger models extract more causal relationships. The rate of increase, however, varies by domain and tends to plateau for larger model sizes, particularly in the deduplicated graphs. The deduplicated graphs (right plot) exhibit a more pronounced plateauing effect and generally lower edge counts than the original graphs (left plot), indicating the effectiveness of the deduplication process in removing redundant information.

of the generated causal graphs. The variation in graph complexity across these domains is illustrated in Figure 3, which compares the average number of edges in the original and deduplicated graphs for each domain.

B. Scaling Behavior

We assess the benchmark tasks using the Qwen 2.5 family of language models, with parameter sizes ranging from 0.5B to 14B. These models are the state-of-the-art series of LLMs developed by Alibaba Cloud. Specifically, we evaluate models ranging from 0.5 billion to 14 billion parameters, allowing us to analyze the impact of model scale on causal knowledge extraction. These models are transformer-based and have been pre-trained on a massive corpus of text and code, equipping them with broad world knowledge and strong language understanding capabilities. Our findings reveal a positive correlation between the number of parameters and the size of the extracted causal graphs, indicating that larger LLMs are capable of identifying more complex causal relationships. This scaling trend is visually evident in Figure 2, which demonstrates the increase in the number of edges in

the generated causal graphs as the model size grows across various scientific domains.

V. CONCLUSION

This work introduces a framework to extract and refine causal knowledge from LLMs' latent representations in the form of Bayesian Networks. We demonstrate that by prompting models to generate causal chains and employing post-processing techniques to merge similar nodes and resolve conflicts, we can construct increasingly complex and accurate causal graphs. Our experiments using the Qwen 2.5 model family show that larger models construct more complex causal explanations.

However, LLMs can produce plausible yet factually inaccurate explanations and may exhibit biases derived from their training data. These biases are particularly pronounced in specialized domains. Addressing these challenges requires a multi-faceted approach. Future work will focus on analyzing and mitigating these biases, potentially by integrating domain-specific knowledge bases to verify generated explanations against established facts. We will also explore

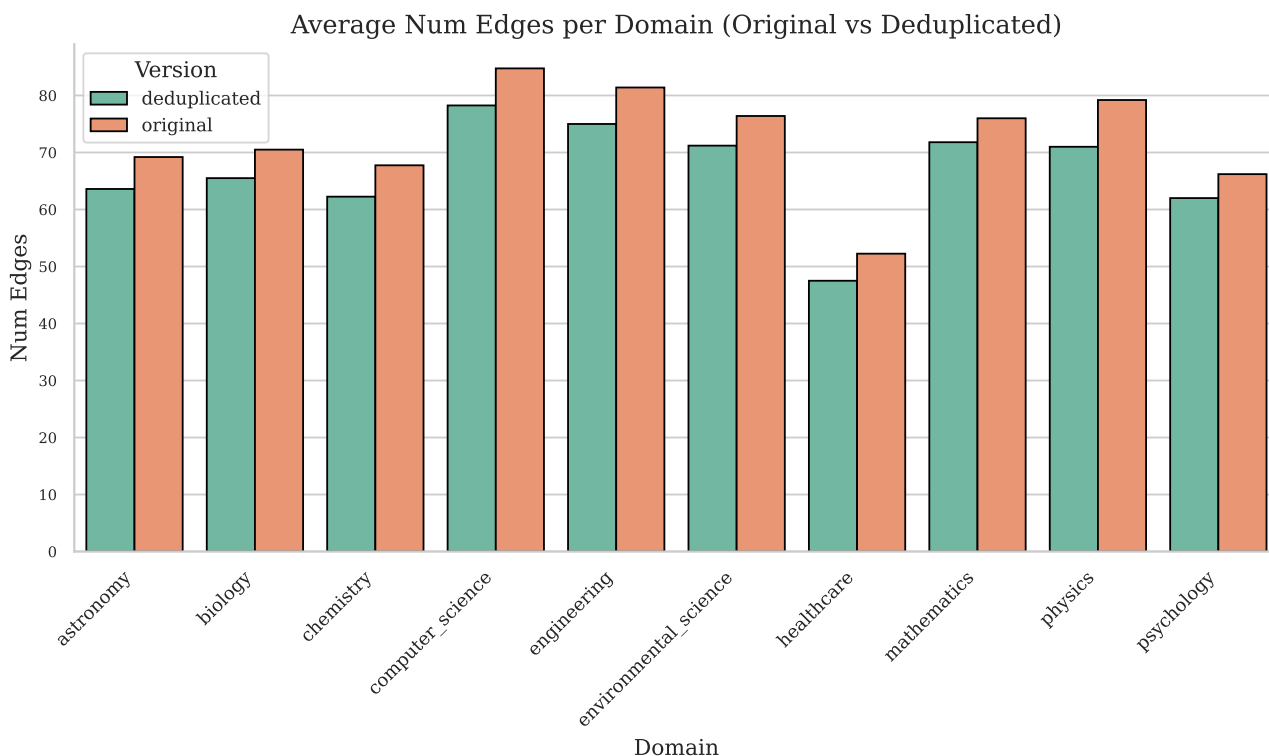


Fig. 3. Average number of edges per domain in causal graphs before and after deduplication. The bar chart compares the average number of edges in causal graphs generated from various scientific domains, contrasting the original graphs with their deduplicated versions. The "original" version (coral bars) represents the initial graph constructed from the extracted causal chains, while the "deduplicated" version (teal bars) reflects the graph after removing redundant edges.

more advanced natural language understanding techniques to handle complex statements involving multiple causes and conditional relationships. One limitation of the current framework is its inability to capture complex contextual relationships where multiple variables jointly influence an outcome under specific conditions. This is due to the chain-based extraction method, which primarily focuses on pairwise causal links. Future work will explore incorporating more sophisticated natural language understanding techniques capable of identifying and representing such context-dependent relationships, potentially by allowing for more complex graph structures beyond simple chains during the initial extraction phase. For example, we could query the LLMs specifically about how different factors interact to produce certain effects or prompt them to describe scenarios where a cause-effect relationship holds only under certain conditions.

REFERENCES

- [1] Peter Antal, Geert Fannes, Dirk Timmerman, Yves Moreau, and Bart De Moor. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in medicine*, 30(3):257–281, 2004.
- [2] Péter Antal and András Millinghofer. Learning causal bayesian networks from literature data. *Periodica Polytechnica Electrical Engineering (Archives)*, 50(3-4):201–221, 2006.
- [3] Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling Causal Reasoning in Large Language Models: Reality or Mirage? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [4] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243, 1995.
- [5] Robert Hoffmann and Alfonso Valencia. Life cycles of successful genes. *Trends in genetics*, 19(2):79–81, 2003.
- [6] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrin-

- maya Sachan, et al. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh conference on neural information processing systems*, 2023.
- [7] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.
- [8] Michael Krauthammer, Charles A Kaufmann, T Conrad Gilliam, and Andrey Rzhetsky. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer’s disease. *Proceedings of the National Academy of Sciences*, 101(42):15148–15153, 2004.
- [9] Michael Krauthammer, Pauline Kra, Ivan Iossifov, Shawn M Gomez, George Hripcsak, Vasileios Hatzivassiloglou, Carol Friedman, and Andrey Rzhetsky. Of truth and pathways: chasing bits of information through myriads of articles. In *ISMB*, pages 249–257, 2002.
- [10] Jing Ma. Causal inference with large language model: A survey. *arXiv preprint arXiv:2409.09822*, 2024.
- [11] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410, 1995.
- [12] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.
- [13] Don R Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986.
- [14] Don R Swanson and Neil R Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial intelligence*, 91(2):183–203, 1997.
- [15] Guangya Wan, Yuqi Wu, Mengxuan Hu, Zhixuan Chu, and Sheng Li. Bridging causal discovery and large language models: A comprehensive survey of integrative approaches and future directions. *arXiv preprint arXiv:2402.11068*, 2024.
- [16] Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck. A Critical Review of Causal Reasoning Benchmarks for Large Language Models. In *AAAI 2024 Workshop on “Are Large Language Models Simply Causal Parrots?”*, 2024.
- [17] Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, and Shangaoran Quan. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.