DKCD: Domain Knowledge-Enhanced Causal Discovery from Unstructured Data

Xin Li[1] **Jin Li**[1] **Shoujin Wang**[1]*
Kun Yu[1] **Fang Chen**[1]
[1] Data Science Institute, University of Technology Sydney, Australia
{xin.li-19, jin.li-4}@student.uts.edu.au,
{shoujin.wang, kun.yu, fang.chen}@uts.edu.au

Abstract

Causal discovery from unstructured data is a challenging yet underexplored task in high-expertise domains such as healthcare, finance, and education. Existing methods typically leverage the general knowledge of large language models (LLMs) to identify causal factors from unstructured data and annotate them into structured data for causal graph construction. However, they remain limited by two key challenges (CHs): **(CH1) insufficient identification of latent factors**, which are implicit in the data yet essential for causal discovery, due to the lack of domain-specific knowledge; and **(CH2) unreliable factor annotation**, caused by the lack of domain-grounded reasoning, which propagates errors to the resulting causal graphs. To address these challenges, we introduce a novel **D**omain **K**nowledge-enhanced **C**ausal **D**iscovery framework (DKCD) for causal discovery from unstructured data in high-expertise domains with three interconnected components: **(1) Knowledge Mining:** It retrieves relevant domain knowledge based on observable factors to support subsequent causal reasoning. **(2) Knowledge-guided Causal Reasoning:** Reasoning with relevant knowledge, it discovers latent causal factors to address CH1 and generates key causal clues for more accurate data annotation to address CH2. and **(3) Causal Structure Discovery:** It constructs the final causal graphs based on a more complete factor set and accurate annotations. Experiments on two domain-specific datasets show that DKCD significantly improves both causal factor identification and causal graph construction. The code and data are available at https://github.com/bigoldli329/DKCD.

1 Introduction

Exploring and uncovering causality from real-world data is fundamental across a wide range of scientific domains (Bunge, 2017; Illari et al., 2011), including healthcare (Shi and Norgeot, 2022; Yang et al., 2013), finance (Papana et al., 2017; Sokolov et al., 2025), and education (Morrison and van der Werf, 2016; Fancsali, 2014). In data-driven causal discovery, causal relationships are typically represented as causal graphs. Existing causal discovery algorithms, such as the Peter–Clark (PC) algorithm (Spirtes et al., 2000; Ramsey, 2016) and the Fast Causal Inference (FCI) algorithm (Spirtes et al., 2013), are primarily designed for structured data with predefined schemes and fixed formats, commonly organized in tabular form (e.g., relational databases and spreadsheets). Thus, they face substantial limitations when applied to unstructured data (Malinsky and Danks, 2018; Yu et al., 2016). Meanwhile, the growing dominance of unstructured data, particularly natural-language text (Siddiqa et al., 2017; Eberendu et al., 2016; Azad et al., 2020), calls for methods that bridge unstructured data and statistical causal discovery.

With the emergence of large language models (LLMs), new opportunities have arisen for analyzing and discovering causal relationships from unstructured data (Wang et al., 2023; Wang, 2024). With strong capabilities for understanding, reasoning, and extracting knowledge from natural language, LLMs can support automatic identification of causal factors, reducing reliance on manual domain expertise for data labeling (Ashwani et al., 2024; Dubey et al., 2024). Specifically, the COAT framework (Liu et al., 2024) is a representative approach that leverages LLMs to extract causal factors from unstructured data, annotate factor values, and then apply statistical causal discovery algorithms to construct causal graphs. Although COAT takes an important step forward, it suffers from critical limitations when applied to high-expertise domains, such as clinical text analysis, where causal factor identification requires substantial domain knowledge, and many causal factors are implicit rather than explicitly stated in unstructured data.

*Corresponding author: shoujin.wang@uts.edu.au

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2824–2850
November 6-10, 2026 ©2026 Association for Computational Linguistics

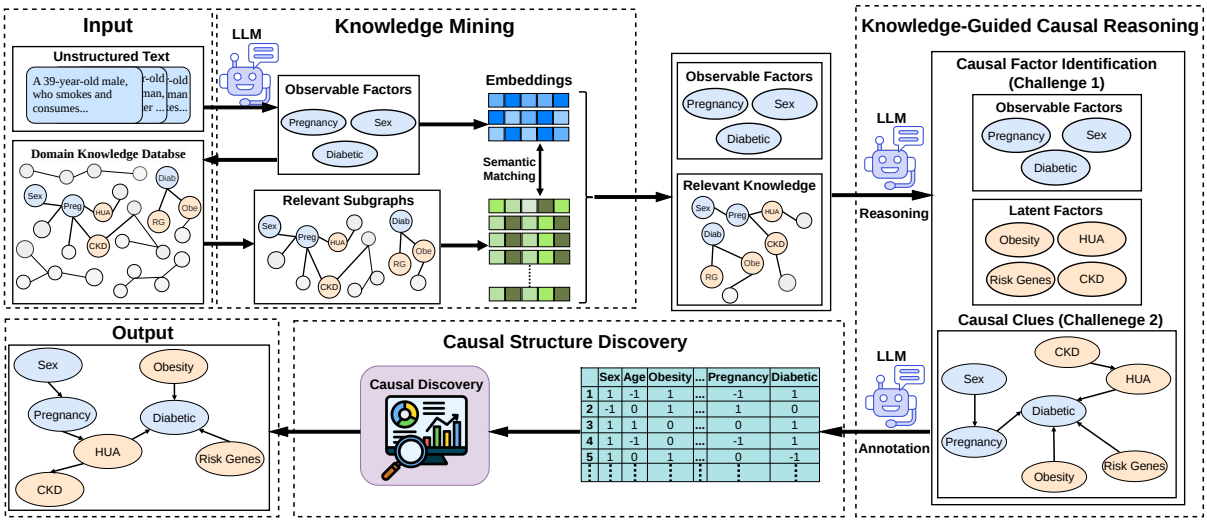


Figure 1: Overview of the DKCD framework. DKCD leverages domain knowledge graphs to guide latent factor discovery and enhances LLM-based causal reasoning by injecting external domain knowledge. The resulting knowledge-grounded factor scoring improves the robustness and accuracy of downstream causal graph construction.

As shown in Figure 2, applying COAT to a diabetes-domain dataset reveals limitations from two key challenges (CHs). **Challenge 1 (CH1):** COAT relies solely on LLMs' general knowledge to propose causal factors. Although such knowledge is broad, it often lacks domain knowledge (Yu et al., 2024b; Shou et al., 2023; Li et al., 2025d), making it difficult to identify latent factors implied by domain-specific observations. For instance, COAT misses kidney disease, highlighted as a red node, which requires knowledge linking reduced kidney function, albuminuria, and diabetes-related complications to renal conditions. Such omissions yield an incomplete factor set and limit the final causal graph. **Challenge 2 (CH2):** Causal graph construction also requires accurate factor-value annotations as structured input to downstream statistical causal discovery algorithms. However, without external domain knowledge, LLMs' limited causal reasoning capabilities (Mirzadeh et al., 2024) often produce inaccurate or inconsistent annotations (Burford et al., 2024; Zečević et al., 2023; Ban et al., 2025). For example, COAT may annotate HUA mainly based on explicit factors such as alcohol or smoking while missing the role of kidney disease, causing errors that propagate into the inferred final causal graph as missing causal relationships and uncertain or incorrect edge directions. Although Li et al. (Li et al., 2025a) extended COAT (Liu et al., 2024) to multimodal datasets through MLLM-CD, the same challenges remain.

To address these challenges, we propose the DKCD (**D**omain **K**nowledge-enhanced **C**ausal **D**iscovery) framework, which effectively leverages

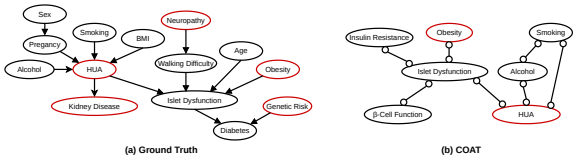


Figure 2: The COAT result on the diabetes dataset.

domain knowledge graphs (KGs) to guide both causal factor discovery and annotation, thereby improving the quality of statistically inferred causal graphs from unstructured data. As illustrated in Figure 1, DKCD consists of three components. (1) Knowledge Mining: It extracts observable factors from unstructured data using LLMs and retrieves relevant subgraphs from domain KGs. Semantic matching filters the retrieved knowledge to preserve the most relevant context, supporting subsequent causal reasoning. (2) Knowledge-Guided Causal Reasoning: Building on the observable factors and retrieved knowledge context, this module infers latent factors that are not explicitly stated in the unstructured data but are causally relevant in the domain, thereby directly addressing the incomplete factor identification challenge (**CH1**). In addition, the module generates *causal clues*, i.e., domain-grounded descriptions of plausible dependencies among factors. These clues provide structured guidance for scoring and annotation, helping reduce ambiguity and improve annotation reliability, thereby addressing the inaccurate factor annotation challenge (**CH2**). (3) Causal Structure Discovery: With a more complete factor set and more accurate annotations, this module applies a statistical causal discovery algorithm to infer the final causal

graph. In addition, Figure 6 in Appendix illustrates DKCD's causal discovery process using a Diabetes dataset example. The main contributions of our work are briefly summarized as follows:

- We propose DKCD, a novel domain knowledge-enhanced causal discovery framework that integrates domain KGs with LLM-based reasoning to enhance causal discovery from unstructured data in high-expertise domains.
- We design a knowledge-guided causal reasoning module that leverages retrieved domain knowledge to discover latent causal factors and generate knowledge-grounded causal clues, thereby improving annotation accuracy.
- We construct the first domain-specific benchmarks with real-world knowledge graphs for evaluation. Extensive experiments show that DKCD significantly outperforms existing methods regarding both causal factor identification and causal graph construction.

2 Related Work

Causal Discovery aims to identify causal relationships from data and typically represents them as causal graphs. Existing methods are commonly divided into three categories: constraint-based (Niu et al., 2024; Vowels et al., 2022), score-based (Vowels et al., 2022), and continuous optimization methods (Gong et al., 2024; Hyttinen et al., 2016). Constraint-based methods infer structure through conditional independence tests, with representative algorithms including FCI (Spirtes et al., 2013), PC (Spirtes et al., 2000; Ramsey, 2016), and PCMCI (Runge et al., 2019). Score-based methods evaluate candidate graphs using criteria such as the Bayesian Information Criterion (BIC) and the Bayesian Dirichlet equivalent uniform (BDeu) (Huang et al., 2018; Nogueira et al., 2022), while continuous optimization methods reformulate structure learning as differentiable optimization, such as NOTEARS (Zhou and Chen, 2022; Lopez et al., 2022). Although existing causal discovery algorithms perform well on structured data, they are difficult to apply directly to unstructured text. DKCD bridges this gap by enabling effective causal discovery on domain-specific unstructured data.

LLM for Causal Discovery has recently emerged as a promising direction (Gopalakrishnan et al., 2024; Wu et al., 2023; Tong et al., 2024). Prior studies use LLMs to extract causal factors (Liu et al., 2024), estimate causal edges (Ban et al., 2023; Darvariu et al., 2024), and support counterfactual reasoning from natural language data (Gendron et al., 2024; Liu et al., 2026). However, LLMs alone do not reliably perform causal reasoning, as their predictions are largely driven by pretraining data, which limits robustness in domain-specific settings.

KG-Guided LLMs for Causal Discovery introduce structured domain knowledge to enhance LLM-based causal reasoning and causal analysis (Kim et al., 2024; Pan et al., 2024). Representative studies include RC2R (Yu et al., 2024a), which focuses on causal reasoning over existing knowledge graphs, and RealTCD (Li et al., 2024b), which performs causal discovery on structured time-series data, using textual information only as auxiliary knowledge. Other studies incorporate biomedical knowledge graphs into prompting or retrieval frameworks to support LLM reasoning (Susanti and Färber, 2024; Lin et al., 2026). Despite these advances, LLMs still struggle to reliably distinguish correlation from causation (Wu et al., 2024), particularly in specialized domains where spurious correlations are common. Moreover, these are not directly applicable to our setting, which requires domain knowledge to extract potential causal factors and reliable annotations from unstructured data.

3 Domain Knowledge-Enhanced Causal Discovery

3.1 Problem Definition

We consider a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, where each sample $\mathbf{x}_k \in \mathcal{X}$ represents a domain-specific textual description. Examples of the generated patient condition descriptions used in our experiments are provided in Appendix B.1. In domain-specific settings such as healthcare, textual descriptions often contain rich unstructured observations that can support the identification of potential causal factors and relationships (Li et al., 2025c; Ban et al., 2025). Let $\mathcal{V}^* = \{V_1, V_2, \ldots, V_d\}$ denote the complete set of ground-truth causal factors underlying the domain process. The ground-truth causal relationships among these variables are represented by a directed acyclic graph (DAG) $G^* = (\mathcal{V}^*, E^*)$, where E^* denotes the set of directed causal edges. Therefore, causal discovery from unstructured data aims to infer a causal graph from $\mathcal{D}$ that approximates the ground-truth structure G^*.

Identifying Causal Factors in Specialized Domains. This task aims to identify a relevant set

of causal factors $\mathcal{V}$ from an unstructured dataset $\mathcal{D}$. However, unstructured data often provides only partial observations of the underlying causal factors and omits several important causal factors, either because they require domain expertise to infer or because they are implicitly reflected through related symptoms, conditions, or measurements. To better formulate this problem, we divide the causal factors to be identified as $\mathcal{V} = \mathcal{V}_o \cup \mathcal{V}_l$. Here, $\mathcal{V}_o$ denotes *observable causal factors*, which are directly mentioned in the observational data and can therefore be easily extracted by existing methods like COAT. $\mathcal{V}_l$ denotes *latent causal factors*, which are not explicitly stated but essential to obtain the correct and comprehensive causal graph. For example, given diabetic patient descriptions, factors such as neuropathy, hyperuricemia (HUA), kidney disease, obesity, and genetic predisposition may not be explicitly mentioned, but they are clinically important for diabetes diagnosis. They can be inferred with domain-knowledge-grounded reasoning from indirect clinical indicators, such as kidney disease, which may be reflected by abnormal eGFR or albuminuria. However, existing methods commonly overlook them (e.g., Figure 2) due to the lack of capabilities to obtain and leverage domain-specific knowledge. Therefore, this motivates the knowledge-guided design of DKCD.

Annotating Factor Values. Given the identified factors $\mathcal{V}$, each text $\mathbf{x}_k$ must be annotated with a value v_{kj} for each variable $V_j \in \mathcal{V}$, converting $\mathcal{D}$ into a structured scoring table

$$\mathbf{S} = \{(v_{k1}, v_{k2}, \cdots, v_{k|\mathcal{V}|})\}_{k=1}^{n}, \tag{1}$$

where $|\mathcal{V}|$ denotes the cardinality of $\mathcal{V}$. Existing methods (Li et al., 2025a; Liu et al., 2024) typically obtain v_{kj} by directly prompting LLMs to annotate the value of each factor for each sample. This annotation step is critical because the quality of $\mathbf{S}$ directly determines the accuracy of downstream causal discovery. However, accurately scoring factors, especially latent ones, in unstructured data requires domain-grounded reasoning, which LLMs lack when operating without external knowledge.

Discovering the Causal Structure. Given the structured data $\mathbf{S}$, a statistical causal discovery algorithm is applied to infer the causal graph G. The accuracy of the inferred graph depends on both the completeness of the factor set and the reliability of the annotations. Errors in either distort conditional independence tests and propagate to the graph.

The proposed DKCD framework aims to address these challenging tasks by integrating LLMs with domain knowledge graphs and statistical causal discovery algorithms for reliable causal structure discovery. The overall framework is illustrated in Figure 1 and consists of three key components. First, the knowledge mining module extracts observable causal factors from unstructured text using LLMs and retrieves relevant subgraphs from the domain knowledge graph to support the subsequent causal reasoning. Second, the knowledge-guided causal reasoning leverages knowledge to discover latent causal factors and derive causal clues, thereby supporting more reliable annotation. Third, the causal structure discovery module applies statistical causal discovery algorithms to infer causal relationships among variables and construct the final causal graph. In addition, Figure 6 provides an illustrative example from the Diabetes dataset to demonstrate the overall process of the DKCD, which is further summarized in Algorithm A.

3.2 Knowledge Mining

As illustrated in Figure 1, this module takes a domain-specific textual description and the domain KG, denoted by $\mathcal{G}_{KG}$, as input. The objective of this module is not only to extract directly observable causal variables, but also to retrieve the most relevant contextual knowledge for subsequent knowledge-guided reasoning.

Observable Factor Proposal: Given a small set of textual samples $\mathcal{X}_s = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, where $\mathcal{X}_s \subset \mathcal{D}$ and each sample $\mathbf{x}_k \in \mathcal{X}_s$ denotes a textual description, the LLM is prompted (see Prompt B.3 in Appendix B.2) to identify a set of observable causal factors $\mathcal{V}_o$ explicitly mentioned in the unstructured data. This step is formulated as the following factor extraction mapping:

$$f_{\text{obs}} : \mathcal{X}_s \rightarrow \mathcal{V}_o. \tag{2}$$

These factors provide the initial semantic anchors for subsequent retrieval of relevant knowledge and latent factor mining. However, these identifying factors alone may still miss important latent causal factors, motivating the use of domain KGs for latent causal factor identification.

Domain KGs Retrieval: Based on the identified observable factors $\mathcal{V}_o$, we retrieve relevant subgraphs from the domain KG, denoted as $\mathcal{G}_{KG}$. Specifically, the observable factors are first normalized without altering their semantic meaning, then matched to relevant entity nodes in $\mathcal{G}_{KG}$. Their neighboring entities and relations are subsequently

collected to construct contextual subgraphs

$$\mathcal{G}_{rel}^{(n)} \subseteq \mathcal{G}_{KG}, \tag{3}$$

where $\mathcal{G}_{rel}^{(n)}$ denotes the retrieved subgraph with n triples, where n is the number of retrieved triples. In the semantic matching stage, these triples are further transformed into natural language statements for relevance ranking. The retrieved subgraph contains structured domain concepts relevant to the observable factors. The domain KG provides external knowledge to enrich textual causal semantics and support latent factor identification.

Semantic Matching: To perform finer-grained retrieval aligned with semantic information, we further conduct semantic matching between observable factors and retrieved knowledge entities to preserve the most relevant knowledge context. Since pretrained sentence encoders are trained on natural language rather than symbolic triples, the retrieved triples are verbalized into natural language statements for semantic similarity matching (Galal et al., 2024; Yin and Zhang, 2024). Specifically, let $\mathcal{S}^{(k)} = \{s_1, s_2, \ldots, s_{n_k}\}$ denote the set of natural language statements verbalized from the triples in the retrieved subgraph. This transformation aligns the structured KG information with the textual embedding space used by pretrained sentence encoders, enabling more effective semantic matching with patient condition descriptions. For each observable factor $v_i \in \mathcal{V}_o$, we encode v_i and each statement $s_j \in \mathcal{S}^{(k)}$ into dense embeddings using the pretrained sentence-transformer model all-MiniLM-L6-v2 (Yin and Zhang, 2024; Vergou et al., 2023), denoted as $E(\cdot)$, and compute their semantic similarity using cosine similarity:

$$\mathrm{sim}(v_i, s_j) = \frac{E(v_i)^\top E(s_j)}{\|E(v_i)\|\|E(s_j)\|}. \tag{4}$$

We then select the top-r most relevant statements for each observable factor:

$$\mathcal{S}_{v_i}^{(k)} = \text{Top-}r\{s_j \in \mathcal{S}^{(k)} \mid \mathrm{sim}(v_i, s_j)\}, \tag{5}$$

where $\mathcal{S}_{v_i}^{(k)}$ denotes the set of matched knowledge statements associated with factor v_i. Further analysis of the parameter r is provided in Appendix B.8. The selected knowledge items are then used to construct the relevant knowledge context for each sample, providing more reliable support for latent factor discovery and causal clue inference.

3.3 Knowledge-Guided Causal Reasoning

This module leverages the retrieved domain knowledge to perform two key reasoning tasks that directly address the two identified challenges. First, it infers latent causal factors absent from the surface text, thereby addressing the problem of incomplete factor identification. Second, it generates knowledge-grounded causal clues that guide more reliable and accurate factor annotation.

Latent Factor Discovery: Given the observable factor set $\mathcal{V}_o$ and the matched knowledge context derived from $\mathcal{G}_{rel}^{(n)}$, the LLM performs knowledge-guided reasoning to infer a set of latent causal factors $\mathcal{V}_l$. These latent factors are not directly mentioned in the unstructured data, but are supported by the semantic relations and domain concepts provided by the retrieved knowledge. The complete causal factor set is therefore defined as

$$\mathcal{V} = \mathcal{V}_o \cup \mathcal{V}_l. \tag{6}$$

By incorporating domain knowledge to discover latent factors, this step extends causal factor discovery beyond surface-level textual mentions, leading to a more comprehensive factor set that better approximates the ground-truth factor set $\mathcal{V}^*$.

Causal Clue Generation: Beyond factor discovery, the LLM derives causal clues from the interaction between the identified factors and the retrieved domain knowledge through internal reasoning ability. These causal clues describe plausible causal dependencies and domain-informed influence patterns among factors. Formally, for each sample $\mathbf{x}_k$, the generated causal clues are denoted as

$$\mathcal{H}^{(k)} = L(\mathcal{V}, \mathcal{G}_{rel}^{(k)}), \tag{7}$$

where $L(\cdot)$ represents the knowledge-guided reasoning process implemented by the LLM following the prompt B.5 shown in Appendix B.2. An example of the derived causal clues is shown in Appendix B.10. Note that the generated clues do not directly determine the final causal graph. Instead, they serve as domain-grounded guidance during the annotation step: by providing the LLM with explicit reasoning about how factors relate causally, the clues help produce more accurate and consistent factor values in the scoring table $\mathbf{S}$, thereby improving the reliability of the structured data fed into the downstream causal structure discovery module.

3.4 Causal Structure Discovery

With a more complete causal factor set $\mathcal{V}$ and more reliable annotations in the structured scoring table $\mathbf{S}$, the final step applies a statistical causal discovery algorithm $\mathcal{C}$ to infer the causal graph. Formally, the causal discovery process can be expressed as

$$G = \mathcal{C}(\mathbf{S}, \mathcal{V}), \tag{8}$$

Table 1: Causal factor identification and structure discovery performance on the diabetes dataset.

LLM	Method	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
GPT-4o	Zero-shot LLM	0.26 ± 0.12	0.14 ± 0.07	0.15 ± 0.03	0.17 ± 0.29	0.03 ± 0.05	0.05 ± 0.08	36.67 ± 2.31
	META	0.57 ± 0.07	0.26 ± 0.11	0.35 ± 0.12	**1.00 ± 0.00**	0.13 ± 0.09	0.22 ± 0.13	32.00 ± 1.00
	COAT	0.61 ± 0.11	0.31 ± 0.09	0.41 ± 0.10	0.65 ± 0.04	0.18 ± 0.04	0.28 ± 0.05	30.67 ± 2.08
	DKCD	**0.84 ± 0.16**	**0.71 ± 0.08**	**0.77 ± 0.04**	0.50 ± 0.17	**0.33 ± 0.18**	**0.40 ± 0.18**	**25.33 ± 3.21**
Gemini 2.5 Flash	Zero-shot LLM	0.35 ± 0.07	0.17 ± 0.04	0.23 ± 0.06	0.50 ± 0.00	0.08 ± 0.00	0.14 ± 0.00	36.00 ± 3.61
	META	0.59 ± 0.08	0.31 ± 0.09	0.34 ± 0.09	0.81 ± 0.17	0.23 ± 0.08	0.36 ± 0.11	**29.67 ± 1.53**
	COAT	0.52 ± 0.25	0.24 ± 0.05	0.32 ± 0.09	0.37 ± 0.32	0.10 ± 0.12	0.16 ± 0.17	35.00 ± 7.00
	DKCD	**0.67 ± 0.03**	**0.62 ± 0.25**	**0.59 ± 0.21**	**0.86 ± 0.15**	**0.33 ± 0.16**	**0.45 ± 0.17**	**29.67 ± 1.53**
Grok-3	Zero-shot LLM	0.54 ± 0.27	0.16 ± 0.04	0.23 ± 0.06	0.17 ± 0.29	0.03 ± 0.05	0.05 ± 0.08	37.33 ± 2.08
	META	**0.78 ± 0.13**	0.41 ± 0.08	0.54 ± 0.09	0.81 ± 0.17	0.28 ± 0.05	0.41 ± 0.06	**27.00 ± 4.58**
	COAT	0.55 ± 0.04	0.26 ± 0.05	0.35 ± 0.05	**0.83 ± 0.29**	0.10 ± 0.04	0.18 ± 0.07	33.00 ± 0.00
	DKCD	0.71 ± 0.04	**0.64 ± 0.07**	**0.67 ± 0.05**	0.74 ± 0.14	**0.41 ± 0.16**	**0.44 ± 0.12**	30.33 ± 3.21
LLaMA 3-70B	Zero-shot LLM	0.56 ± 0.22	0.38 ± 0.18	0.45 ± 0.20	0.20 ± 0.09	0.29 ± 0.09	0.29 ± 0.09	29.67 ± 8.14
	META	**0.79 ± 0.10**	0.41 ± 0.04	0.53 ± 0.04	0.48 ± 0.14	0.20 ± 0.05	0.29 ± 0.07	39.00 ± 3.46
	COAT	0.74 ± 0.07	0.38 ± 0.08	0.50 ± 0.07	0.60 ± 0.35	0.18 ± 0.05	0.27 ± 0.09	29.00 ± 3.61
	DKCD	0.71 ± 0.06	**0.57 ± 0.00**	**0.63 ± 0.07**	**0.65 ± 0.19**	**0.33 ± 0.12**	**0.44 ± 0.14**	**27.67 ± 3.51**

where $\mathcal{C}$ denotes the causal discovery algorithm, and G denotes the inferred causal graph whose nodes correspond to the variables in $\mathcal{V}$ and whose edges represent the discovered causal relationships. Moreover, we instantiate $\mathcal{C}$ using the FCI algorithm (Spirtes et al., 2013) for final causal structure discovery. FCI is particularly suitable because it can infer causal structures in the presence of latent confounders without requiring all causal factors to be observed (Ramsey et al., 2025; Varghese et al., 2024), which is important in specialized domains where some causal factors may remain unidentified. Since traditional causal discovery algorithms, including FCI, require structured tabular data and are sensitive to noisy or incomplete inputs (Malinsky and Danks, 2018), the upstream modules of DKCD are designed to produce a more complete factor set and more accurate annotations, helping recover a causal graph closer to the ground truth G^*.

4 Experiments

We evaluate DKCD for causal factor and structure discovery on synthetic Diabetes and Respiratory datasets with GPT-4o (Hurst et al., 2024), Gemini 2.5 Flash (Comanici et al., 2025), Grok-3 (xAI, 2025), and LLaMA 3-70B (MetaAI, 2024). Appendices B.1–B.5 describe dataset/KG construction, prompts, implementation, metrics, and baselines. Appendix B.6 reports full results, including latent-factor recovery (LFR), retrieval-augmented generation (RAG), and GPT-5.5 (OpenAI, 2026) comparisons; Appendices B.7 and B.8 provide complete ablation and parameter analyses. Appendix B.9 assesses robustness across FCI and PC, sensitivity to KG incompleteness and noise, and potential data leakage from LLM memorization of the public KG.

4.1 Experimental Setup

Currently, there are no publicly available datasets that simultaneously provide unstructured data and aligned ground-truth causal graphs in domain-specific settings. Since causal discovery requires ground-truth causal graphs for quantitative evaluation, synthetic datasets remain necessary in existing research practice (Liu et al., 2024; Li et al., 2025a). Following prior work, we construct two medical-domain synthetic datasets grounded in real-world medical KGs for evaluation: (1) the **Diabetic Patient Condition Descriptions Dataset (Diabetes Dataset)**, which contains 400 patient condition descriptions with 14 high-level factors and a diabetes-domain KG; and (2) the **Respiratory Patient Condition Descriptions Dataset (Respiratory Dataset)**, which contains 400 patient condition descriptions with 8 high-level factors and a respiratory-domain KG. Additional details are provided in Appendix B.3.

Diabetes Dataset: Following prior work (Liu et al., 2024; Li et al., 2025a), we generate a medical dataset containing clinical descriptions of diabetic patients. Each synthetic sample contains a textual description of patient conditions and clinical information. This includes 9 observable factors (i.e., sex, pregnancy, alcohol, smoking, BMI, walking difficulty, age, diabetes, and islet dysfunction), and 5 latent factors (HUA, kidney disease, neuropathy, obesity, and genetic risk), as shown in Figure 3(a). Additionally, we construct a diabetes-domain KG using the DiaKG dataset (Chang et al., 2021). Additional construction details and the example of the generated data are provided in Appendix B.1.1.

Table 2: Causal factor identification and structure discovery performance on the respiratory dataset.

LLM	Method	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
GPT-4o	Zero-shot LLM	0.42 ± 0.07	0.38 ± 0.00	0.40 ± 0.03	0.44 ± 0.10	0.13 ± 0.00	0.20 ± 0.01	20.00 ± 5.20
	META	0.33 ± 0.06	0.25 ± 0.00	0.28 ± 0.02	**1.00 ± 0.00**	0.13 ± 0.00	0.23 ± 0.00	25.67 ± 1.53
	COAT	0.34 ± 0.06	0.25 ± 0.00	0.29 ± 0.02	0.67 ± 0.58	0.09 ± 0.08	0.15 ± 0.13	27.00 ± 4.00
	DKCD	**0.74 ± 0.13**	**0.75 ± 0.00**	**0.74 ± 0.07**	0.87 ± 0.23	**0.56 ± 0.22**	**0.68 ± 0.21**	**16.67 ± 2.31**
Gemini 2.5 Flash	Zero-shot LLM	0.39 ± 0.05	0.29 ± 0.08	0.33 ± 0.06	**0.44 ± 0.51**	0.09 ± 0.08	0.14 ± 0.12	23.00 ± 6.24
	META	0.21 ± 0.06	0.13 ± 0.00	0.16 ± 0.02	0.33 ± 0.58	0.04 ± 0.08	0.08 ± 0.13	26.00 ± 4.36
	COAT	0.23 ± 0.06	0.21 ± 0.07	0.22 ± 0.07	0.33 ± 0.58	0.04 ± 0.08	0.08 ± 0.13	30.00 ± 1.73
	DKCD	**0.84 ± 0.08**	**0.88 ± 0.13**	**0.86 ± 0.10**	0.41 ± 0.16	**0.42 ± 0.19**	**0.41 ± 0.17**	**17.33 ± 3.21**
Grok-3	Zero-shot LLM	0.51 ± 0.09	0.38 ± 0.00	0.41 ± 0.02	0.44 ± 0.10	0.13 ± 0.00	0.20 ± 0.01	22.33 ± 3.21
	META	0.37 ± 0.04	0.29 ± 0.08	0.33 ± 0.05	**1.00 ± 0.00**	0.13 ± 0.00	0.23 ± 0.00	25.67 ± 2.08
	COAT	0.32 ± 0.18	0.21 ± 0.07	0.25 ± 0.11	0.33 ± 0.58	0.04 ± 0.08	0.08 ± 0.13	26.33 ± 6.03
	DKCD	**0.67 ± 0.14**	**0.63 ± 0.21**	**0.64 ± 0.18**	0.62 ± 0.13	**0.34 ± 0.08**	**0.43 ± 0.09**	**18.67 ± 1.15**
LLaMA 3-70B	Zero-shot LLM	0.32 ± 0.02	0.38 ± 0.00	0.34 ± 0.01	0.50 ± 0.00	0.13 ± 0.00	0.21 ± 0.00	27.00 ± 2.00
	META	0.35 ± 0.07	0.29 ± 0.08	0.32 ± 0.07	**1.00 ± 0.00**	0.13 ± 0.00	0.21 ± 0.00	25.67 ± 0.58
	COAT	0.53 ± 0.12	0.42 ± 0.07	0.47 ± 0.09	0.83 ± 0.29	0.25 ± 0.00	0.38 ± 0.04	23.33 ± 3.06
	DKCD	**0.73 ± 0.09**	**0.80 ± 0.14**	**0.69 ± 0.11**	0.81 ± 0.17	**0.50 ± 0.25**	**0.57 ± 0.18**	**17.33 ± 5.03**

Respiratory Dataset: We also generate another medical dataset containing clinical descriptions of patients with respiratory conditions, following prior work (Liu et al., 2024; Li et al., 2025a). Each sample includes 5 observable factors (i.e., tuberculosis, smoke, lung cancer, lung disease, and chest X-ray) and 3 latent factors (i.e., Asian travel, bronchitis, and dyspnea), as illustrated in Figure 11(a). Additionally, we construct a respiratory-domain KG based on prior studies. Additional construction details and the example of the generated data are provided in Appendix B.1.2.

Baselines: Since causal discovery directly from unstructured data remains underexplored, we include all relevant and comparable baselines in the literature. Zero-shot LLM (Antonucci et al., 2023; Du et al., 2025) generates causal graphs directly using LLMs without domain knowledge. COAT (Liu et al., 2024) is a state-of-the-art LLM-driven framework that combines LLM with traditional causal discovery algorithms. META (Liu et al., 2024) is a variant derived from the COAT ablation study. Other related methods are not included because they focus on different settings, such as causal reasoning over existing KGs (Yu et al., 2024a) or causal discovery from structured data (Li et al., 2024b), rather than directly discovering causal graphs from unstructured data. Additional details are provided in Appendix B.5.

Metrics: Following common practice in the causal discovery literature, we evaluate the performance of factor identification and graph structure using widely adopted metrics (Zanga et al., 2022; Li et al., 2025a; Liu et al., 2024). For causal factor identification, we report: (1) Node Precision (NP), (2) Node Recall (NR), and (3) Node F1-score (NF). For causal structure discovery, we adopt: (1) Adjacency Precision (AP), (2) Adjacency Recall (AR) and (3) Adjacency F1-score (AF). To jointly assess both aspects, we also employ an extended Structural Hamming Distance (ESHD) (Peters and Bühlmann, 2015; Li et al., 2025a). All results are reported as the mean and standard deviation over three runs. Additional details are provided in Appendix B.4.

4.2 Analysis on the Datasets

The empirical results of causal discovery on the diabetes and respiratory datasets are presented in Table 1 and Figure 3, and Table 2 and Figure 11, respectively. The results demonstrate that DKCD consistently outperforms the baseline methods in most cases while maintaining stable performance across different LLM backbones.

Results on the Diabetes Dataset: DKCD achieves better overall causal factor identification performance across most LLM settings, obtaining higher NP, NR, and NF scores than the baselines, as shown in Table 1. Compared with COAT (Liu et al., 2024), DKCD more accurately identifies both observable and latent causal factors, showing the effectiveness of incorporating domain KGs and latent factor reasoning for causal discovery from unstructured data. DKCD also achieves competitive graph structure performance, with lower ESHD values and stable AR and AF scores across different LLM backbones, indicating closer alignment with the ground-truth causal graph, as illustrated in Figure 3. Additional graph comparison results across different LLM backbones are provided in Appendix B.6.

Results on the Respiratory Dataset: A similar trend can be observed on the respiratory dataset.

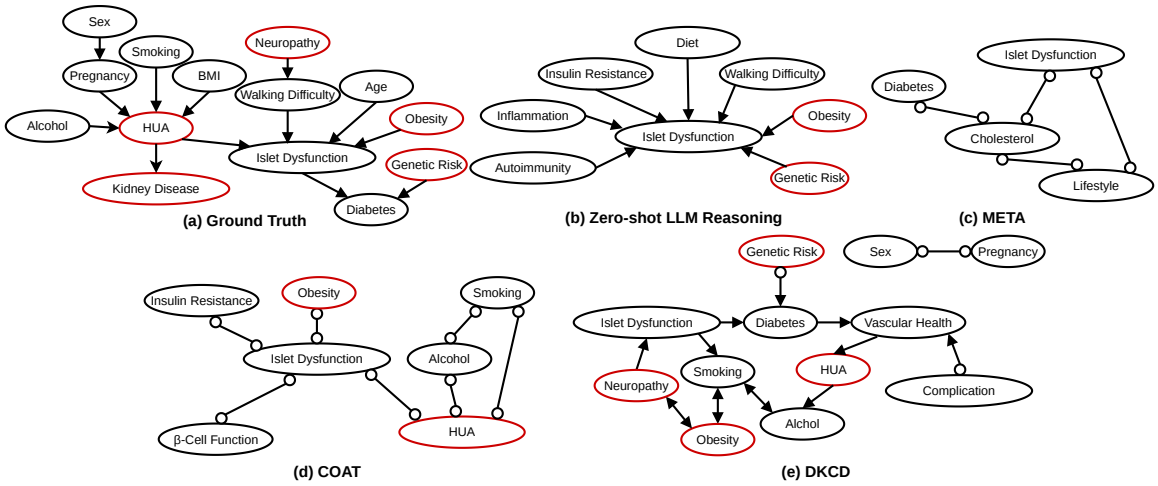


Figure 3: Ground-truth and discovered causal graphs on the diabetes dataset. Latent factors are marked in red.

In terms of causal factor identification, DKCD achieves the highest NP, NR, and NF scores across most LLM configurations, with quantitative results presented in Table 2. These results indicate that DKCD has a strong ability to extract meaningful causal factors from patient descriptions of respiratory conditions. In terms of causal graph structure generation, DKCD also achieves lower ESHD values and higher AR and AF scores in most cases. These results indicate that DKCD can effectively recover the underlying causal structure among respiratory-related factors. Overall, the results further confirm that integrating domain knowledge with causal discovery algorithms significantly improves causal discovery from unstructured data. Additional graph comparison results across different LLM backbones are provided in Appendix B.6.

4.3 Ablation Study

We conduct an ablation study using GPT-4o as the representative LLM backbone to evaluate the effectiveness of key components in DKCD. Specifically, we compare the full DKCD model with three variants: (1) w/o CC removes causal clues in the knowledge-guided causal reasoning module; (2) w/o LF removes latent factor support, causing the annotation process to rely solely on LLM-identified factors; and (3) w/o Both removes both causal clues and latent factor identification, simulating the setting without knowledge graphs. Additional ablation results and analyses of different design choices are provided in Appendix B.7.

The experimental results on the Diabetes and Respiratory datasets are shown in Table 3. Removing causal clues mainly degrades structural metrics such as AF and ESHD, while the NF score remains unchanged because causal clues are introduced after factor mining and therefore do not affect node identification. In contrast, removing latent factor identification leads to larger performance drops, especially in NF, highlighting the importance of domain knowledge for discovering causal latent factors. Removing both components causes the severe degradation, indicating that the two designs are complementary. Overall, DKCD achieves competitive performance across both datasets, demonstrating the benefit of integrating domain knowledge with reasoning guidance for causal discovery.

Table 3: Ablation study of the DKCD framework on the Diabetes and Respiratory datasets.

Dataset	Variant	NF↑	AF↑	ESHD↓
Diabetes dataset	w/o CC	**0.77 ± 0.04**	0.38 ± 0.15	26.67 ± 2.52
	w/o LF	0.33 ± 0.14	0.00 ± 0.00	26.67 ± 2.52
	w/o Both	0.54 ± 0.31	0.10 ± 0.09	34.00 ± 1.73
	DKCD	**0.77 ± 0.04**	**0.40 ± 0.18**	**25.33 ± 3.21**
Respiratory dataset	w/o CC	**0.74 ± 0.07**	0.62 ± 0.10	17.00 ± 2.65
	w/o LF	0.45 ± 0.11	0.32 ± 0.16	22.33 ± 4.51
	w/o Both	0.33 ± 0.15	0.14 ± 0.12	25.33 ± 1.53
	DKCD	**0.74 ± 0.07**	**0.68 ± 0.21**	**16.67 ± 3.21**

5 Conclusions

This paper presents DKCD, a knowledge-enhanced framework for causal discovery from unstructured data in high-expertise domains. DKCD consists of three key components: a knowledge mining module that retrieves relevant domain knowledge, a knowledge-guided causal reasoning module that discovers latent causal factors and generates key causal clues for accurate annotation, and a causal structure discovery module that constructs the final causal graph from the annotation table. Extensive experiments on two domain datasets show that DKCD significantly outperforms existing methods.

Limitations

Although DKCD integrates LLMs with domain knowledge graphs to improve causal discovery from unstructured data, several limitations remain.

Reliance on LLM-generated reasoning. DKCD still relies on LLMs' reasoning and generation capabilities. Although our design provides domain-knowledge-grounded reasoning, LLMs may still generate hallucinated, inconsistent, or biased causal contents, especially in complex domain-specific scenarios (Gu et al., 2026; Shi et al., 2025; Yang et al., 2025a).

Limited availability of benchmark datasets. Causal discovery from unstructured domain-specific data is limited by the lack of publicly available benchmark datasets containing both unstructured textual descriptions and aligned ground-truth causal graphs. Constructing reliable benchmarks in specific domains remains challenging because it requires extensive domain expertise, causal annotation, and validation (Visbal et al., 2004; Rodrigues et al., 2022). Although the datasets used in this work are carefully constructed based on prior literature and grounded in real-world medical knowledge, they should be viewed as a starting point for developing and evaluating methods in this emerging direction of domain-specific unstructured causal discovery. Developing larger-scale and more diverse benchmarks will be a focus of future work.

Scalability to broader domains. Our experiments focus on two medical-domain datasets, which allow us to evaluate DKCD in specialized settings where domain knowledge is important (Yang et al., 2026a,c, 2025b). However, the generalization of DKCD to other domains, such as education, finance, and the social sciences, remains underexplored (Yang et al., 2026b).

Future work will address these limitations in three main directions. First, we will develop stronger validation mechanisms for LLM-generated causal reasoning, including consistency checking, uncertainty estimation, and causal constraints derived from domain knowledge. With the rapid development of diffusion models in multiple domains (Ho et al., 2020; Cai et al., 2026a, 2025, 2026c), some work explores applying diffusion models to conduct causal reasoning (Duan et al., 2024; Lin et al., 2025; Li et al., 2025b). This technique can also be utilized to improve the reliability of causal reasoning. Second, we will expand the evaluation of DKCD by constructing and testing on more diverse datasets, including real-world domain-specific textual data, larger-scale knowledge graphs, and expert-validated causal structures. Third, we will investigate the application of DKCD to broader specific domains, such as discovering causal factors encoded in misinformation to improve the content credibility and trustworthiness of recommender systems (Cai et al., 2026b; Li et al., 2024a, 2026).

Ethical Considerations

This research focuses on methodological work for constructing causal graphs from unstructured textual data. The datasets used in our experiments are either synthetic or derived from publicly available sources, and do not contain personally identifiable information. Additionally, our framework is intended to assist researchers in discovering potential causal structure and should not be used as a substitute for professional judgment, particularly in sensitive domains such as healthcare. Future applications should carefully consider issues related to data privacy, the potential biases inherent in language models, and the responsible generation, interpretation, and use of automatically inferred causal relationships in practical settings.

Acknowledgments

This research program relies on student participation, which was made possible through student funding provided under the Next Generation Artificial Intelligence Graduates Program (NGAIGP), conducted by the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO).

References

Alessandro Antonucci, Gregorio Piqué, and Marco Zaffalon. 2023. Zero-shot causal graph extrapolation from text via llms. *arXiv preprint arXiv:2312.14670*.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Dushyant Singh Sengar, Mayank Jindal, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, pages 2–9.

Poopak Azad, Nima Jafari Navimipour, Amir Masoud Rahmani, and Arash Sharifi. 2020. The role of structured and unstructured data managing mechanisms in the internet of things. *Cluster Computing*, 23(2):1185–1198.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. 2023. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, Qiang Tu, and Huanhuan Chen. 2025. Integrating large language model for improved causal discovery. *IEEE Transactions on Artificial Intelligence*.

Mario Bunge. 2017. *Causality and modern science*. Routledge.

Kelly G. Burford, Nathan G. Itzkowitz, Anthony G. Ortega, Julien O. Teitler, and Andrew G. Rundle. 2024. Use of generative ai to identify helmet status among patients with micromobility-related injuries from unstructured clinical notes. *Jama Network Open*, 7(8):e2425981. 10.1001/jamanetworkopen.2024.25981.

Zhuo Cai, Shoujin Wang, Victor W Chu, Usman Naseem, Yang Wang, and Fang Chen. 2025. Unleashing the potential of diffusion models towards diversified sequential recommendations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1476–1486.

Zhuo Cai, Shoujin Wang, Jin Li, Peilin Zhou, Victor W Chu, Fang Chen, Tianqing Zhu, and Charu Aggarwal. 2026a. Steering diffusion models towards credible content recommendation. In *International Conference on Learning Representations*, volume 2026, pages 42132–42163.

Zhuo Cai, Shoujin Wang, Peilin Zhou, Yan Wang, Xiao Liu, Lianyong Qi, Julian McAuley, and Dietmar Jannach. 2026b. Misinformation unlearning for responsible content recommendation. *ACM Transactions on Information Systems*, 44(5):1–40.

Zhuo Cai, Shoujin Wang, Peilin Zhou, Min Xu, Julian McAuley, and Fang Chen. 2026c. Sprint: Single-step generative recommendation via average probability velocity. *arXiv preprint arXiv:2609.34306*.

Maxwell Caplin and Freda Festenstein. 1975. Relation between lung cancer, chronic bronchitis, and airways obstruction. *British Medical Journal*, 3(5985):678–680.

Dejie Chang, Mosha Chen, Chaozhen Liu, Liping Liu, Dongdong Li, Wei Li, Fei Kong, Bangchang Liu, Xiaobin Luo, Ji Qi, and 1 others. 2021. Diakg: An annotated diabetes dataset for medical knowledge graph construction. In *China Conference on Knowledge Graph and Semantic Computing*, pages 308–314. Springer.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. 2024. Large language models are effective priors for causal graph discovery. *arXiv preprint arXiv:2405.13551*.

Dario Di Palma, Felice Antonio Merra, Maurizio Sfilio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. Do llms memorize recommendation datasets? a preliminary study on movielens-1m. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2582–2586.

Huaming Du, Yujia Zheng, Baoyu Jing, Yu Zhao, Gang Kou, Guisong Liu, Tao Gu, Weimin Li, and Carl Yang. 2025. Causal discovery through synergizing large language model and data-driven reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 543–554.

Yifan Duan, Jian Zhao, Junyuan Mao, Hao Wu, Jingyu Xu, Shilong Wang, Caoyuan Ma, Kai Wang, Kun Wang, and Xuelong Li. 2024. Causal deciphering and inpainting in spatio-temporal dynamics via diffusion model. *Advances in Neural Information Processing Systems*, 37:107604–107632.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Adanma Cecilia Eberendu and 1 others. 2016. Unstructured data: an overview of the data of big data. *International Journal of Computer Trends and Technology*, 38(1):46–50.

Stephen Fancsali. 2014. Causal discovery with models: behavior, affect, and learning in cognitive tutor algebra. In *Educational Data Mining 2014*.

Omar Galal, Ahmed H Abdel-Gawad, and Mona Farouk. 2024. Rethinking of bert sentence embedding for text classification. *Neural Computing and Applications*, 36(32):20245–20258.

Gaël Gendron, Jože M Rožanec, Michael Witbrock, and Gillian Dobbie. 2024. Counterfactual causal inference in natural language with large language models. *arXiv preprint arXiv:2410.06392*.

Chang Gong, Chuzhe Zhang, Di Yao, Jingping Bi, Wenbin Li, and Yongjun Xu. 2024. Causal discovery from temporal data: An overview and new perspectives. *ACM Computing Surveys*, 57(4):1–38.

Seethalakshmi Gopalakrishnan, Luciana Garbayo, and 1 others. 2024. Causality extraction from medical text using large language models (llms). *CoRR abs/2407.10020*.

Huilin Gu, Shoujin Wang, Jin Li, Zhuo Cai, Lina Yao, Yan wang, and Longbing Cao. 2026. A survey on trustworthy llm-based recommender systems.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. 2018. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. 2016. Causal discovery from subsampled time series data by constraint optimization. In *Conference on Probabilistic Graphical Models*, pages 216–227. PMLR.

Phyllis McKay Illari, Federica Russo, and Jon Williamson. 2011. *Causality in the Sciences*. Oxford University Press.

Yejin Kim, Eojin Kang, Juae Kim, and H Howie Huang. 2024. Causal reasoning in large language models: A knowledge graph approach. *arXiv preprint arXiv:2410.11588*.

Steffen L Lauritzen and David J Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194.

Jin Li, Huilin Gu, Shoujin Wang, Qi Zhang, Shui Yu, Chen Wang, Xiwei Xu, and Fang Chen. 2026. Towards fair large language model-based recommender systems without costly retraining. In *Proceedings of the ACM Web Conference*, pages 9675–9686.

Jin Li, Shoujin Wang, Qi Zhang, Longbing Cao, Fang Chen, Xiuzhen Zhang, Dietmar Jannach, and Charu C. Aggarwal. 2024a. Causal learning for trustworthy recommender systems: A survey. *CoRR*, abs/2402.08241.

Jin Li, Shoujin Wang, Qi Zhang, Feng Liu, Tongliang Liu, Longbing Cao, Shui Yu, and Fang Chen. 2025a. Revealing multimodal causality with large language models. *arXiv preprint arXiv:2509.17784*.

Jin Li, Shoujin Wang, Qi Zhang, Shui Yu, and Fang Chen. 2025b. Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations. In *Proceedings of the ACM on Web Conference*, pages 2787–2798.

Junyi Li, Yongqiang Chen, Chenxi Liu, Qianyi Cai, Tongliang Liu, Bo Han, Kun Zhang, and Hui Xiong. 2025c. Can large language models help experimental design for causal discovery? *arXiv preprint arXiv:2503.01139*.

Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenwu Zhu. 2024b. Realtcd: Temporal causal discovery from interventional data with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4669–4677.

Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. 2025d. A survey on enhancing causal reasoning ability of large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 399–416. Springer.

Damon Lin, Marzieh Mussavi Rizi, Conor O'Neill, Jeffrey C Lotz, Paul Anderson, and Abel Torres Espin. 2026. Knowledge augmented causal discovery through large language models and knowledge graphs: application in chronic low back pain. *MedRxiv*, pages 2026–02.

Haohong Lin, Xin Huang, Tung Phan, David Hayden, Huan Zhang, Ding Zhao, Siddhartha Srinivasa, Eric Wolff, and Hongge Chen. 2025. Causal composition diffusion model for closed-loop traffic generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27542–27552. IEEE.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. 2024. Discovery of the hidden world with large language models. *Advances in Neural Information Processing Systems*, 37:102307–102365.

Zhonghua Liu, Wei Shao, Xia Cao, and 1 others. 2026. Crest: A causal framework for mitigating shortcut learning in language models through counterfactual reasoning. *Information Processing & Management*, 63(2):104418.

Romain Lopez, Jan-Christian Hütter, Jonathan Pritchard, and Aviv Regev. 2022. Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35:19290–19303.

Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470.

Abraham S Meijnikman, Victor E Gerdes, Max Nieuwdorp, and Hilde Herrema. 2018. Evaluating causality of gut microbiota in obesity and diabetes in humans. *Endocrine reviews*, 39(2):133–153.

MetaAI. 2024. Introducing meta llama 3: The most capable openly available llm to date. Ai.meta.com/blog/meta-llama-3/.

Justin J Miller. 2013. Graph database applications and concepts with neo4j. In *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*, volume 2324, pages 141–147. AISEL.

Iman Mirzadeh, Keivan Alizadeh, and 1 others. 2024. Gsm-symbolic: understanding the limitations of mathematical reasoning in large language models. *CoRR abs/2410.05229*.

Esther Molina-Montes, Claudia Coscia, Paulina Gómez-Rubio, Alba Fernández, Rianne Boenink, Marta Rava, Mirari Márquez, Xavier Molero, Matthias Löhr, Linda Sharp, and 1 others. 2021. Deciphering the complex interplay between pancreatic cancer, diabetes mellitus subtypes and obesity/bmi through causal inference and mediation analyses. *Gut*, 70(2):319–329.

Keith Morrison and Greetje van der Werf. 2016. Searching for causality in educational research.

Wenjin Niu, Zijun Gao, Liyan Song, and Lingbo Li. 2024. Comprehensive review and empirical evaluation of causal discovery algorithms for numerical data. *arXiv preprint arXiv:2407.13054*.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. 2022. Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2):e1449.

Mi Jin Noh and Yang Sok Kim. 2025. Diabetes prediction through linkage of causal discovery and inference model with machine learning models. *Biomedicines*, 13(1):124.

Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. 2012. Hamming distance metric learning. *Advances in Neural Information Processing Systems*, 25.

Ikushi Onozaki, Irwin Law, Charalambos Sismanidis, Matteo Zignol, Philippe Glaziou, and Katherine Floyd. 2015. National tuberculosis prevalence surveys in asia, 1990–2012: an overview of results and lessons learned. *Tropical Medicine & International Health*, 20(9):1128–1145.

OpenAI. 2026. Introducing GPT-5.5: A new class of intelligence for real work.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Angeliki Papana, Catherine Kyrtsou, Dimitris Kugiumtzis, and Cees Diks. 2017. Financial networks based on granger causality: a case study. *Physica A: Statistical Mechanics and its Applications*, 482:65–73.

Jonas Peters and Peter Bühlmann. 2015. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799.

Jay Pujara, Hui Miao, Lise Getoor, and William Cohen. 2013. Knowledge graph identification. In *International Semantic Web Conference*, pages 542–557. Springer.

Humairat H Rahman, Danielle Niemann, and Stuart H Munson-McGee. 2023. Association between asthma, chronic bronchitis, emphysema, chronic obstructive pulmonary disease, and lung cancer in the us population. *Environmental Science and Pollution Research*, 30(8):20147–20158.

Joseph Ramsey. 2016. Improving accuracy and scalability of the pc algorithm by maximizing p-value. *arXiv preprint arXiv:1610.00378*.

Joseph Ramsey, Bryan Andrews, and Peter Spirtes. 2025. Efficient latent variable causal discovery: Combining score search and targeted testing. *arXiv preprint arXiv:2510.04263*.

Daniela Rodrigues, Noemi Kreif, Anna Lawrence-Jones, Mauricio Barahona, and Erik Mayer. 2022. Reflection on modern methods: constructing directed acyclic graphs (dags) with domain experts for health services research. *International Journal of Epidemiology*, 51(4):1339–1348.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996.

Jingpu Shi and Beau Norgeot. 2022. Learning causal effects from observational data in healthcare: a review and summary. *Frontiers in Medicine*, 9:864882.

Yuhong Shi, Kun Yu, Yifei Dong, and Fang Chen. 2025. Large language models in education: a systematic review of empirical applications, benefits, and challenges. *Computers and Education: Artificial Intelligence*, page 100529.

Xiao Shou, Debarun Bhattacharjya, Tian Gao, Dharmashankar Subramanian, Oktie Hassanzadeh, and Kristin P Bennett. 2023. Pairwise causality guided transformers for event sequences. *Advances in Neural Information Processing Systems*, 36:46520–46533.

Aisha Siddiqa, Ahmad Karim, and Abdullah Gani. 2017. Big data storage technologies: a survey. *Frontiers of Information Technology & Electronic Engineering*, 18(8):1040–1070.

Alik Sokolov, Fabrizzio Sabelli, Behzad Azadie Faraz, Wuding Li, and Luis Seco. 2025. Toward automating causal discovery in financial markets and beyond. In *Transactions of ADIA Lab: Interdisciplinary Advances in Data and Computational Science*, pages 359–392. World Scientific.

Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. MIT Press.

Peter L Spirtes, Christopher Meek, and Thomas S Richardson. 2013. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*.

Percy Stocks. 1959. Cancer and bronchitis mortality in relation to atmospheric deposit and smoke. *British Medical Journal*, 1(5114):74.

Yuni Susanti and Michael Färber. 2024. Knowledge graph structure as prompt: Improving small language models capabilities for knowledge-based causal discovery. In *International Semantic Web Conference*, pages 87–106. Springer.

Song Tong, Kai Mao, and 1 others. 2024. Automating psychological hypothesis generation with ai: large language models meet causal graph. *CoRR abs/2402.14424*.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. 2006. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.

Akhil Varghese, Miguel Arana-Catania, Stefano Mori, Adriana Encinas-Oropesa, and Joy Sumner. 2024. Causal discovery to understand hot corrosion. *Materials and Corrosion*, 75(12):1639–1651.

Elena Vergou, Ioanna Pagouni, Marios Nanos, and Katia Lida Kermanidis. 2023. Readability classification with wikipedia data and all-minilm embeddings. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 369–380. Springer.

Antonio L Visbal, Brent A Williams, Francis C Nichols III, Randolph S Marks, James R Jett, Marie-Christine Aubry, Eric S Edell, Jason A Wampfler, Julian R Molina, and Ping Yang. 2004. Gender differences in non–small-cell lung cancer survival: an analysis of 4,618 patients diagnosed between 1997 and 2002. *The Annals of thoracic surgery*, 78(1):209–215.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. 2022. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. A causal view of entity bias in (large) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184.

Zeyu Wang. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319*.

Shijie Wu, Ozan Irsoy, and 1 others. 2023. Bloomberggpt: a large language model for finance. *CoRR abs/2303.17564*.

xAI. 2025. Grok 3 beta — the age of reasoning agents. X.ai/news/grok-3.

Jing Yang, Ning An, Gil Alterovitz, Lian Li, and Aiguo Wang. 2013. Causal discovery based on healthcare information. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 71–73. IEEE.

Xiaoyu Yang, Jie Lu, and En Yu. 2025a. Adapting multimodal large language model to concept drift from pre-training onwards. In *International Conference on Learning Representations*, pages 90869–90891.

Xiaoyu Yang, Jie Lu, and En Yu. 2026a. Walking the tightrope: Autonomous disentangling beneficial and detrimental drifts in non-stationary custom-tuning. *Advances in neural information processing systems*, 38:116167–116193.

Xiaoyu Yang, Lijian Xu, Hongsheng Li, and Shaoting Zhang. 2025b. One leaf reveals the season: Occlusion-based contrastive learning with semantic-aware views for efficient visual representation. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 71425–71440.

Xiaoyu Yang, Lijian Xu, Xingyu Zeng, Xiaosong Wang, Hongsheng Li, and Shaoting Zhang. 2026b. Scalar: Spatial-concept alignment for robust vision in harsh open world. *Pattern Recognition*, page 113203.

Xiaoyu Yang, En Yu, Wei Duan, and Jie Lu. 2026c. Turning drift into constraint: Robust reasoning alignment in non-stationary multi-stream environments. In *Forty-third International Conference on Machine Learning*.

Chen Yin and Zixuan Zhang. 2024. A study of sentence similarity based on the all-minilm-l6-v2 model with "same semantics, different structure" after fine tuning. In *2024 2nd International Conference on Image, Algorithms and Artificial Intelligence*, pages 677–684. Atlantis Press.

Guanyuan Yu, Xv Wang, Qing Li, and Yu Zhao. 2024a. Fusing llms and kgs for formal causal reasoning behind financial risk contagion. *arXiv preprint arXiv:2407.17190*.

Kui Yu, Jiuyong Li, and Lin Liu. 2016. A review on algorithms for constraint-based causal discovery. *arXiv preprint arXiv:1611.03977*.

Xuanqing Yu, Wangtao Sun, Jingwei Li, Kang Liu, Chengbao Liu, and Jie Tan. 2024b. Onsep: a novel online neural-symbolic framework for event prediction based on large language model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6335–6350.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. 2024. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8.

Wenxiu Zhou and QingCai Chen. 2022. A survey on causal discovery. In *China Conference on Knowledge Graph and Semantic Computing*, pages 123–135. Springer.

Xiaohan Zou. 2020. A survey on application of knowledge graph. In *Journal of Physics: Conference Series*, volume 1487, page 012016. IOP Publishing.

A Algorithm

Algorithm 1: The DKCD Framework

Input: $\mathcal{D}$, $\mathcal{X}_s$, $\mathcal{G}_{KG}$, f_{obs}, M, L, $\mathcal{C}$.
Output: $\mathcal{V}$, $\mathbf{S}$, G.

1. Propose observable factors: $\mathcal{V}_o \leftarrow f_{\text{obs}}(\mathcal{X}_s)$.
2. Initialize $\mathbf{S} \leftarrow [\,]$.
3. For each $\mathbf{x}_k \in \mathcal{D}$:
   (a) Retrieve subgraph: $\mathcal{G}_{rel}^{(k)} \subseteq \mathcal{G}_{KG}$.
   (b) Verbalize subgraph: $\mathcal{S}^{(k)} = \{s_1, \ldots, s_{n_k}\}$.
   (c) Semantic matching: $\mathcal{S}_{v_i}^{(k)} \leftarrow M(v_i, \mathcal{S}^{(k)})$, $\forall v_i \in \mathcal{V}_o$.
   (d) Build matched context: $\widetilde{\mathcal{S}}^{(k)} = \{\mathcal{S}_{v_i}^{(k)} \mid v_i \in \mathcal{V}_o\}$.
   (e) Discover latent factors: $\mathcal{V}_l \leftarrow L(\mathcal{V}_o, \widetilde{\mathcal{S}}^{(k)})$.
   (f) Generate causal clues: $\mathcal{H} \leftarrow L(\mathcal{V}_o \cup \mathcal{V}_l, \mathcal{G}_{rel}^{(k)})$.
   (g) Annotate factors: $\mathcal{V} \leftarrow \mathcal{V}_o \cup \mathcal{V}_l$, $\mathbf{s}^{(k)} \leftarrow (v_{k1}, \ldots, v_{k|\mathcal{V}|})$.
   (h) Update scoring table: $\mathbf{S} \leftarrow \mathbf{S} \cup \{\mathbf{s}^{(k)}\}$.
4. Discover causal graph: $G \leftarrow \mathcal{C}(\mathbf{S}, \mathcal{V})$.
5. Return $\mathcal{V}, \mathbf{S}, G$.

B More Details about Experiments

We construct two medical-domain datasets: the Diabetic Patient Condition Descriptions Dataset (Diabetes Dataset) and the Respiratory Patient Condition Descriptions Dataset (Respiratory Dataset). Each dataset contains 400 patient condition descriptions, a set of high-level factors, and a corresponding domain knowledge graph.

B.1 Dataset Construction Process

B.1.1 Diabetes Dataset

Diabetic Patient Condition Descriptions Generation: We construct a dataset of diabetic patient condition descriptions following the methodology proposed in (Liu et al., 2024). As illustrated in Figure 4, we first construct a ground-truth causal graph based on domain-specific academic materials (Noh and Kim, 2025; Molina-Montes et al., 2021; Meijnikman et al., 2018), and then generate structured data samples in the form of a scoring table according to the predefined causal relationships encoded in the graph. The FCI algorithm is then applied to the generated structured dataset to reconstruct the causal graph for causal consistency verification. Only when the reconstructed graph is structurally consistent with the ground-truth causal graph are the structured samples used to generate descriptions of diabetic patients' conditions. Finally, Prompt B.1 is used to convert the generated structured samples into unstructured diabetic patient condition descriptions.

Example Generated Diabetic Patient Condition Description: An example generated diabetic patient condition description is shown below:

> A middle-aged woman, a non-smoker with no alcohol use history, presents for routine assessment amidst long-standing diabetes and hypertension. She maintains a BMI of 24 kg/m^2, though she experiences activity-related limitations attributed to her weight, potentially suggesting underlying obesity. There is no report of numbness or tingling, inconsistent with neuropathy. Family history reveals no genetic predisposition to diabetes or associated disorders. Vitals: blood pressure is well-controlled at 128/82 mmHg. Laboratory results indicate a fasting plasma glucose of 5.2 mmol/L, 2-hour OGTT of 6.4 mmol/L, and HbA1c at 5.5%, suggesting

normal glucose metabolism. Kidney function tests show an eGFR of 58 mL/min/1.73 m^2, with albuminuria at 32 mg/g, suggesting chronic kidney disease. Lipid panel results are within normal limits, showing LDL-C at 2.2 mmol/L, HDL-C at 1.4 mmol/L, and triglycerides at 1.0 mmol/L. Pathophysiologically, derived indices such as HOMA-IR $\approx$ 1.5 and HOMA-β $\approx$ 80% indicate maintained insulin sensitivity and robust β-cell function. Cardiometabolic risk context is moderate, influenced by her weight considerations and renal findings but offset by optimal lipid levels and normoglycemic status. Diabetes status: absent; metrics do not signify diabetes, sustaining a diagnosis of normoglycemia.

Diabetes-Domain KG Construction: We construct the diabetes-domain KG using Neo4j (Miller, 2013), based on the DiaKG resource provided in (Chang et al., 2021), which contains annotated entities and relations for diabetes-related medical knowledge. Specifically, each annotated entity is converted into a Neo4j node containing its name, type, and source information, while each annotated relation is converted into an edge connecting its head and tail entities. For example, the annotated triplet (Insulin resistance, Pathogenesis_Disease, Obesity) is imported as two entity nodes connected by a Pathogenesis_Disease relation.

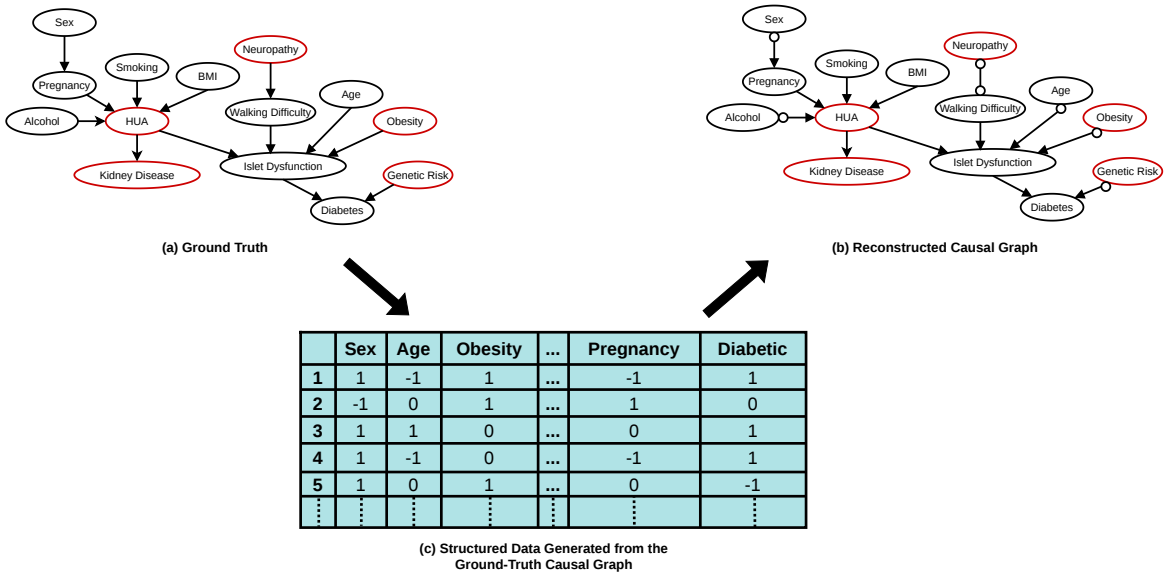
B.1.2 Respiratory Dataset

Respiratory Patient Condition Descriptions Generation: We construct a dataset of respiratory patient condition descriptions following the methodology proposed in (Liu et al., 2024). As illustrated in Figure 5, we first construct a ground-truth causal graph based on prior domain knowledge and the causal structure introduced in (Lauritzen and Spiegelhalter, 1988), and then generate structured data samples in the form of a scoring table according to the predefined causal relationships encoded in the graph. The FCI algorithm is then applied to the generated structured dataset to reconstruct the causal graph for verification of causal consistency. Only when the reconstructed graph is structurally consistent with the ground-truth causal graph are the structured samples used to generate corresponding descriptions of respiratory patient conditions. Finally, Prompt B.2 is used to convert the generated structured samples into unstructured respiratory patient condition descriptions.

Example Generated Respiratory Patient Condition Description: An example generated respiratory patient condition description is shown below:

A 34-year-old woman presents to the clinic for a routine check-up without any specific complaints. She denies recent travel, including visits to Asia, or any known exposure to tuberculosis. The patient is a non-smoker and has no significant respiratory history. She reports being in good health overall, with no history of cough, sputum production, fever, weight loss, night sweats, or hemolysis. Additionally, she does not experience dyspnea, either at rest or with exertion, and her daily activities are not limited. Vital signs are stable, with a temperature of 36.7°C, heart rate 72 bpm, respiratory rate 16/min, blood pressure 118/76 mmHg, and oxygen saturation 98% on room air. Chest examination reveals normal breath sounds with no wheeze or crackles. Routine laboratory tests show a normal white blood cell count and inflammatory markers. A chest X-ray performed as part of her health maintenance check reveals normal findings without evidence of infiltrates or abnormal pulmonary opacities. Clinical impression indicates no underlying lung condition, and the chest X-ray supports this normal finding. She is reassured, and no further pulmonary investigation is needed at this time. Recommendations include maintaining a healthy lifestyle and routine follow-up as part of preventive healthcare.

Respiratory-Domain KG Construction: Since suitable respiratory-domain knowledge graphs are difficult to obtain, we collect a set of publications in the respiratory domain (Rahman et al., 2023; Stocks, 1959; Lauritzen and Spiegelhalter, 1988; Onozaki et al., 2015; Caplin and Festenstein, 1975). Based on these publications, we employ GPT-4o to extract relational triples in the form of (entity, relation, entity), which serve as the basic structure for KG construction (Pujara et al., 2013; Zou, 2020). The extracted nodes and edges, together with their supporting source passages, are imported into Neo4j (Miller, 2013). For example, a source



	Sex	Age	Obesity	...	Pregnancy	Diabetic
1	1	-1	1	...	-1	1
2	-1	0	1	...	1	0
3	1	1	0	...	0	1
4	1	-1	0	...	-1	1
5	1	0	1	...	0	-1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 4: Ground truth causal graph and the causal graph reconstructed by the FCI algorithm from the generated structured data derived from the ground truth on the diabetes dataset. Latent factors are marked in red.

passage describing the association between smoking and lung cancer is converted into the triplet (Smoke, Association, Lung Cancer).

B.2 Details on Prompts

In this section, we provide examples of prompts used in our experiments, including:

- Generating diabetic patient condition descriptions for the diabetes dataset (Prompt B.1).
- Generating respiratory patient condition descriptions for the respiratory dataset (Prompt B.2).
- An example of observable causal factor mining (Prompt B.3).
- An example of knowledge-guided latent causal factor mining (Prompt B.4).
- An example of causal clues generation for causal structure discovery (Prompt B.5).
- An example of the text evaluation prompt (Prompt B.6).

Prompt B.1: Generating Diabetic Patient Condition Descriptions for the Diabetes Dataset

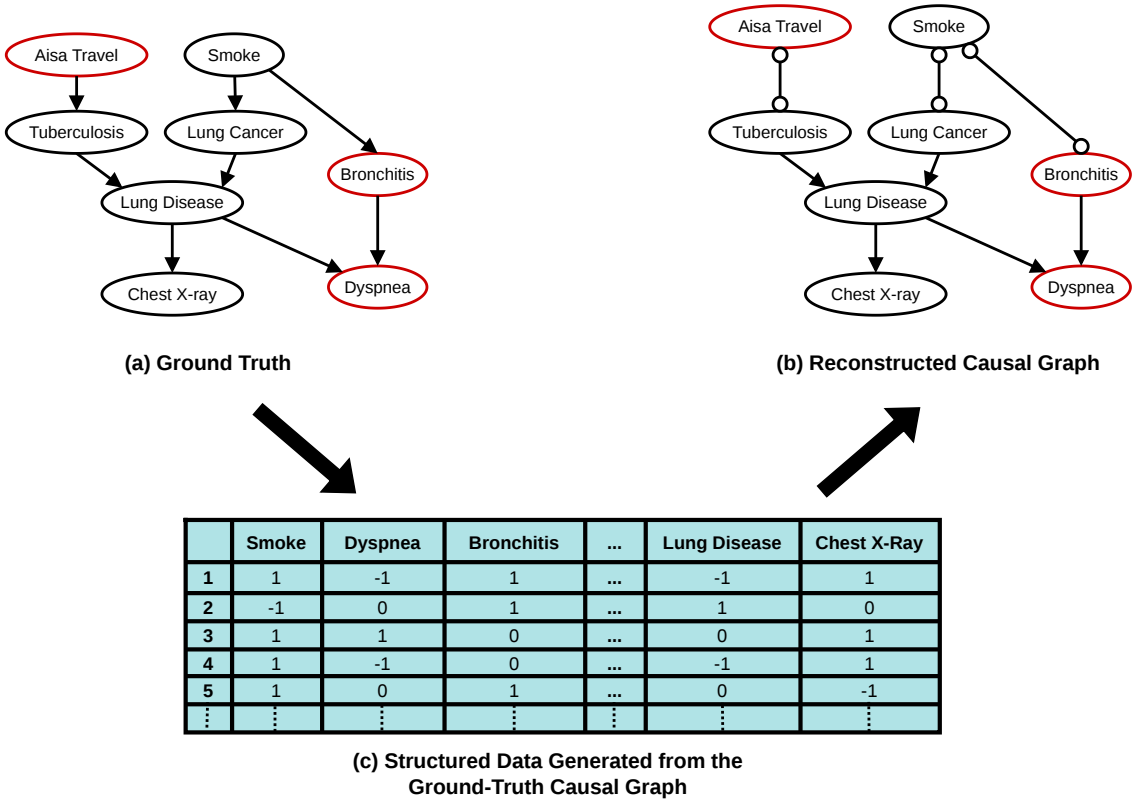
Task:
Please write a clinical description for a patient based on the given diabetic evaluation results.
Evaluation Results: {Diabetic Factors}
Requirement:
- Integrate all provided factors into a clinically plausible patient condition description.
- Embed natural clinical clues that imply the following latent factors without explicitly labeling them:

  - Neuropathy: numbness, tingling, burning sensation, or reduced sensation in the feet.
  - Obesity: elevated body weight, increased BMI, or weight-related mobility limitations.
  - HUA (Hyperuricemia): gout attacks, elevated serum urate, or urate-lowering medication.
  - ChronicKidneyDisease: albuminuria, reduced eGFR, kidney impairment, or long-standing metabolic disease.
  - RiskGenes: family history of diabetes or genetic predisposition to metabolic disorders.

- Single paragraph; No quotation marks; The description should be complete.
- Modern clinical English.
- 150–250 words.
- Only output the patient condition description directly without any other format or content.

Prompt B.2: Generating Respiratory Patient Condition Descriptions for the Respiratory Dataset

Task:
Please write a clinical description for a patient based on the given respiratory evaluation results.
Evaluation Results: {Respiratory Factors}
Requirement:
- Integrate all provided factors into a clinically plausible patient condition description.
- Embed natural clinical clues that imply the following



	Smoke	Dyspnea	Bronchitis	...	Lung Disease	Chest X-Ray
1	1	-1	1	...	-1	1
2	-1	0	1	...	1	0
3	1	1	0	...	0	1
4	1	-1	0	...	-1	1
5	1	0	1	...	0	-1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 5: Ground-truth causal graph, generated structured dataset, and the causal graph reconstructed by the FCI algorithm on the respiratory dataset. The ground-truth graph is first used to generate structured data samples, from which the FCI algorithm reconstructs the causal graph. Latent factors are highlighted in red.

latent factors without explicitly labeling them:

- Visit_to_Asia: travel or exposure history related to TB risk.
- Bronchitis: cough, sputum production, wheezing, or bronchitic symptoms.
- Dyspnea: degree of shortness of breath or breathing difficulty.

- Single paragraph; No quotation marks; The description should be complete.
- Modern clinical English.
- 150–250 words.
- Only output the patient condition description directly without any other format or content.

Prompt B.3: An Example of Observable Causal Factor Mining

You are an expert clinical analyst specializing in medical record interpretation. You are ready to analyze contrasting patient condition descriptions and identify observable causal factors.
Task:
Patient Condition Descriptions: {Patient Condition Descriptions}
Based on these patient condition descriptions, analyze the underlying differences among causal factors that may contribute to the observed distinctions.
Requirements:
- Identify the key observable factors that differentiate the patient records.
- Each factor should focus on one concrete aspect without semantic overlap.
- Keep the criterion for value 0 fixed as: not exist; or not mentioned.
- Each factor is only allowed to take one value from {-1, 0, 1}.
Output Format:
Your output should contain the following three parts.
Part 1: Consideration
Explain your reasoning process when identifying causal factors from the patient records. Focus on how the textual evidence supports the abstraction of candidate factors.
Part 2: Factor Filtration
Evaluate each candidate factor and decide whether to retain it.
Selection causal factor criteria:

- Each factor should help distinguish different patient conditions.
- Each factor should represent a single concrete aspect.
- Semantic overlap between factors should be minimized.

Part 3: Final Output
Report the final factors using the following template:
{Factor Name}

- - 1: [Positive Criterion]
- - 0: [Not exist; or not mentioned]
- - -1: [Negative Criterion]

Prompt B.4: An Example of Knowledge-guided Latent Causal Factor Mining

Use the patient condition description, relevant domain knowledge, and observable causal factor criteria to identify clinically meaningful **latent causal factors**.
Patient Condition Descriptions: {Patient Condition Descriptions}
Relevant Domain Knowledge: {Relevant Domain Knowledge}
Observable Causal Factor Criteria: {Observable Causal Factor Criteria}
Task:
Based on the patient condition description and the relevant domain knowledge, identify **latent causal factors**. These latent factors should capture deeper causal abstractions rather than directly repeating observable factors.
Requirements

- Latent factors should represent higher-level mechanisms or conceptual causes.
- They must be distinct from observable factors.
- Do not duplicate, paraphrase, or rename observable factors.
- Avoid semantic overlap with observable factors.

Output Format:
Part 1: Consideration
Briefly describe how you abstract latent factors from observable factors and domain knowledge.
Part 2: Factor Filtration
Explain why each candidate factor should be kept or discarded.
Part 3: Final Output
Report the final factors using the following template:
{Factor Name}

- - 1: [Positive Criterion]
- - 0: [Not exist; or not mentioned]
- - -1: [Negative Criterion]

Prompt B.5: An Example of Causal Clues Generation for Causal Structure Discovery

Task:
Using ONLY the factors finalized according to {Patient Condition Description} and {Causal Factor Criteria}, infer causal clues.
Patient Condition Description: {Patient Condition Description}
Causal Factor Criteria: {Causal Factor Criteria}
Relevant Domain Knowledge: {Relevant Domain Knowledge}
Requirements
- Infer causal clues only when there is clear textual evidence from the patient condition description and causal factor criteria.
- Each causal clue should suggest a potential directional relationship between two factors.
- If the relationship is uncertain or unsupported by the text, omit the edge rather than guessing.
Output Format
Part 1: Inferred Causal Clues
List the inferred causal clues in the following format:
Source_Factor → Target_Factor
Part 2: Justification
Provide a brief justification for each clue based on the patient's condition description.

Prompt B.6: An Example of Text Annotation Prompt

You are a highly capable assistant for analyzing, abstracting, and processing textual data.
Task:
Evaluate the following factors based on the provided patient condition description.
Patient Condition Descriptions: {Patient Condition Descriptions}
Each factor must be assigned exactly one value from {-1, 0, 1} according to its evaluation criteria.
Factor Criteria and Causal Clue (used as reference during scoring):
{Factor Criteria} {Causal Clue}
Requirements
- For each factor, select exactly one value from {-1, 0, 1}.
- If the factor is not mentioned in the text, return 0.
- The assigned value must strictly follow the given factor criteria.
Output Format
Return ONLY a single-line compact JSON object:
Rules:

- Output JSON only (no explanations, no markdown, no extra text).
- Keys must match exactly (including spaces and lowercase).
- Values must be integers from {-1, 0, 1}.

Example format: {json_example}

B.3 Implementation Details

We implement DKCD as follows. In the knowledge mining module, we first sample representative patient condition descriptions from different outcome groups. Following the COAT method, we empirically select up to 20 samples per group to form a diverse, balanced input set for identifying observable factors. For relevant domain knowledge retrieval, we normalize the expressions of the extracted observable factors without changing their meanings and use fuzzy matching to locate the corresponding Neo4j entities. We then retrieve their one-hop neighbors and connecting relations to ob-

tain candidate subgraphs. The retrieved triplets are verbalized into knowledge sentences and ranked against the observable-factor descriptions using embeddings from the pretrained sentence-transformer all-MiniLM-L6-v2 (Yin and Zhang, 2024; Vergou et al., 2023) and cosine similarity. We retain the top-r contexts for each factor. For example, the observable factor Renal Function can retrieve the triplet (Renal function, Test_Disease, Chronic kidney disease), which is verbalized and supplied as relevant knowledge for subsequent latent-factor reasoning. Based on the parameter analysis in Table 12, we set $r = 20$ to balance relevant knowledge coverage and contextual noise.

In the knowledge-guided causal reasoning module, the LLM takes representative patient condition descriptions, observable factor criteria, and retrieved knowledge contexts as input to infer a shared set of latent causal factors at the corpus level. The complete factor set is then annotated separately for every document. Causal clues guide annotation but do not directly determine the final graph. In particular, background knowledge alone is insufficient to assign a positive value: the document must contain direct or indirect supporting evidence. When neither is available, the corresponding value is set to 0 (not observed). In the causal structure discovery module, we apply the FCI algorithm implemented in the causal-learn library (Zheng et al., 2024) to the annotated scoring table to infer the final causal graph structure. All experimental results are reported as the mean and standard deviation over three independent runs. All experiments are conducted on a server with two Intel Xeon 6346 CPUs, 256GB RAM, and two NVIDIA A40 GPUs, with all LLMs accessed via API calls.

B.4 Evaluation Metrics

To evaluate the effectiveness of our method, we adopt a set of commonly used metrics for both causal factor identification and causal graph structure discovery, following prior work (Li et al., 2025a; Liu et al., 2024; Zanga et al., 2022).

Factor Identification: We evaluate the accuracy and completeness of the identified causal factors using precision (NP), recall (NR), and F1 score (NF). Since factor names may vary across different methods and models, we manually align semantically equivalent factors with the ground-truth names during evaluation to ensure consistency.

Latent Factor Recovery: We additionally report the latent factor recovery rate (LFR), defined as

$$\mathrm{LFR} = \frac{N_{\mathrm{recovered}}}{N_{\mathrm{latent}}},$$

where $N_{\mathrm{recovered}}$ is the number of successfully recovered ground-truth latent factors and N_{latent} is the total number of ground-truth latent factors. LFR directly measures the ability of a method to recover factors that are implicit in the documents rather than explicitly observed.

Causal Graph Structure Discovery: The quality of the discovered causal structure is evaluated using adjacency precision (AP), adjacency recall (AR), and their F1 score (AF). These adjacency metrics evaluate whether two factors are connected but do not distinguish between opposite edge directions. We further adopt the extended structural Hamming distance (ESHD) metric (Li et al., 2025a). Compared with the standard SHD metric (Tsamardinos et al., 2006; Norouzi et al., 2012), ESHD additionally accounts for direction errors, missing and spurious factors, and their associated edges.

B.5 Baselines

We provide a detailed list of the representative baselines used in our experiments, including:

- **Zero-shot LLM** (Antonucci et al., 2023; Du et al., 2025) relies solely on the reasoning capability of LLMs without providing any examples or domain knowledge.
- **Zero-shot LLM + RAG** augments the zero-shot LLM with retrieved domain context. Unlike DKCD, it does not include dedicated modules for structured KG retrieval, latent-factor reasoning, causal-clue generation, or evidence-grounded annotation.
- **META** (Liu et al., 2024) is a variant derived from the COAT ablation study, where contextual examples are not provided. It performs zero-shot causal factor proposal based only on the given context using LLMs.
- **COAT** (Liu et al., 2024) is a state-of-the-art LLM-driven framework for causal discovery from unstructured text. It integrates LLM reasoning with traditional causal discovery algorithms to construct causal graphs.

B.6 Full Experimental Results

The complete qualitative comparison of different methods across multiple LLM backbones, in-

Table 4: Latent factor recovery (LFR) rate on the Diabetes dataset using GPT-4o.

Dataset	Method	LFR↑
Diabetes dataset	Zero-shot LLM	0.27 ± 0.12
	META	0.27 ± 0.23
	COAT	0.40 ± 0.20
	DKCD	$\mathbf{0.67 \pm 0.12}$

cluding GPT-4o, Gemini 2.5 Flash, Grok-3, and LLaMA 3-70B, on the Diabetes dataset is presented in Figures 7, 8, 9, and 10. These visual results offer a more comprehensive view of the causal graph structures produced by each method under different LLM backbones, thereby facilitating a detailed qualitative assessment of their structural discovery capabilities. Likewise, the corresponding qualitative comparisons on the Respiratory dataset are presented in Figures 11, 12, 13, and 14.

B.6.1 Latent Factor Recovery

Table 4 reports LFR on the Diabetes dataset using GPT-4o. DKCD achieves the highest LFR of 0.67 ± 0.12, improving over the second-best method, COAT, by 0.27 absolute (27 percentage points). This result is consistent with both the factor-identification metrics and the qualitative comparisons of discovered latent factors.

B.6.2 Additional Baseline Comparisons

We additionally compare DKCD with a retrieval-augmented zero-shot baseline and with zero-shot prompting of a more recent LLM. As shown in Table 5, simply providing retrieved domain information to a zero-shot LLM is insufficient: Zero-shot LLM + RAG obtains an NF of 0.16 ± 0.05 and fails to recover any correct adjacency. The zero-shot GPT-5.5 baseline may identify some relevant factors, but it does not reliably recover a complete factor set or causal structure. In contrast, DKCD combines structured retrieval, latent-factor reasoning, causal-clue generation, and evidence-grounded annotation, yielding substantially better factor and graph recovery.

B.7 Full Ablation Study Results

To complement the findings in Section 4.4, we further present detailed ablation results on the Diabetes and Respiratory datasets using GPT-4o as a representative LLM backbone. These results offer a more fine-grained analysis of each component's role in the proposed framework for causal factor identification and causal structure discovery. Specifically, they allow us to examine how removing individual components affects both factor-level performance and the structural quality of the inferred causal graphs. The complete results are reported in Table 10.

Additionally, in Table 11, we study key design choices in DKCD (Figure 1). SM denotes semantic matching, while SP refers to replacing the structured prompt (Prompt B.4) in knowledge-guided causal reasoning with a simpler task-only prompt that omits the reasoning chain. As shown, removing either component degrades performance, with noticeable drops in both factor identification (NP/NR/NF) and structural metrics (AR/AF/ESHD), highlighting the importance of both semantic matching and structured reasoning.

B.8 Parameter Analysis Results

In this section, we analyze the impact of the parameter r on DKCD performance on the Diabetes dataset using GPT-4o. Here, r denotes the number of top-r relevant knowledge contexts selected for each factor after semantic matching. When r is small, insufficient knowledge is retrieved, leading to lower performance. As shown in Table 12, performance improves as r increases and reaches a relatively good level at around $r = 20$. In contrast, incorporating too much contextual knowledge can introduce excessive noise, thereby degrading DKCD performance. Therefore, we use $r = 20$ in the main experiments.

B.9 Additional Robustness and Sensitivity Analyses

In this section, we conduct additional experiments to evaluate the robustness and reliability of DKCD under different downstream causal discovery methods and varying KG conditions. Specifically, we examine its sensitivity to KG incompleteness and noise and assess whether GPT-4o memorizes the public Diabetes KG.

B.9.1 Robustness to the Causal Discovery Method

To assess whether DKCD depends on the downstream causal discovery algorithm, we replace FCI with PC on the Diabetes dataset using GPT-4o. As reported in Table 6, factor-identification results remain unchanged because both variants use the same factor-identification and annotation stages. PC yields comparable graph-recovery performance,

Table 5: Additional baseline comparisons on the Diabetes dataset.

LLM	Method	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
GPT-4o	Zero-shot LLM + RAG	0.24 ± 0.10	0.12 ± 0.04	0.16 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	36.67 ± 2.31
GPT-5.5	Zero-shot LLM	0.22 ± 0.04	0.19 ± 0.04	0.20 ± 0.04	$\mathbf{0.50 \pm 0.50}$	0.05 ± 0.05	0.10 ± 0.08	54.00 ± 7.00
GPT-4o	**DKCD**	$\mathbf{0.84 \pm 0.16}$	$\mathbf{0.71 \pm 0.08}$	$\mathbf{0.77 \pm 0.04}$	$\mathbf{0.50 \pm 0.17}$	$\mathbf{0.33 \pm 0.18}$	$\mathbf{0.40 \pm 0.18}$	$\mathbf{25.33 \pm 3.21}$

Table 6: Robustness to the downstream causal discovery method on the Diabetes dataset using GPT-4o.

Causal discovery method	Method	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
FCI	DKCD	0.84 ± 0.16	0.71 ± 0.08	0.77 ± 0.04	0.50 ± 0.17	0.33 ± 0.18	0.40 ± 0.18	25.33 ± 3.21
PC	DKCD	0.84 ± 0.16	0.71 ± 0.08	0.77 ± 0.04	$\mathbf{0.56 \pm 0.20}$	$\mathbf{0.39 \pm 0.31}$	$\mathbf{0.43 \pm 0.29}$	$\mathbf{24.33 \pm 2.52}$

Table 7: Exact-match memorization check on the Diabetes KG using GPT-4o.

Metric	Result
Total source sentences	8,643
Unique triples	3,967
Held-out queries	3,921
Exact-match accuracy	0.15

with slightly higher AP, AR, and AF and lower ESHD. Although PC assumes causal sufficiency and FCI allows latent confounding, DKCD remains effective with both methods, indicating that its performance is not specific to FCI.

B.9.2 Sensitivity to KG Incompleteness

We evaluate the influence of KG completeness by randomly removing 10%, 20%, or 30% of the nodes and edges from the Diabetes KG. Table 8 shows that performance generally degrades as more knowledge is removed, especially for factor recall and ESHD. Nevertheless, the results under moderate removal demonstrate that the coarse entity/subgraph retrieval and fine-grained semantic matching stages can still recover useful knowledge from an incomplete KG.

B.9.3 Sensitivity to KG Noise

We further inject synthetic noise into 10%, 20%, or 30% of the Diabetes KG triples. As shown in Table 9, limited noise can be partially filtered by semantic matching and evidence-constrained LLM reasoning, while larger noise ratios substantially reduce factor recall and graph quality. Retrieved knowledge is used as guidance for reasoning and causal-clue generation rather than being inserted directly into the discovered graph, which helps DKCD tolerate moderate corruption. The apparently high AP at the 30% noise level should not be interpreted as a more accurate graph: AP ignores edge direction and can be high when only a small number of adjacencies are predicted, whereas ESHD also penalizes direction errors and missing or spurious factors and edges.

B.9.4 Potential KG Data Leakage from LLM Memorization

Because the Diabetes KG is derived from a publicly available resource, its content may have been included in GPT-4o's training data, potentially introducing data leakage through memorization. Following the exact-match memorization evaluation in (Di Palma et al., 2025), we assess whether GPT-4o can reproduce the KG triples directly from its parametric memory. We construct 3,921 held-out queries from 3,967 unique triples derived from 8,643 source sentences. As shown in Table 7, GPT-4o achieves an exact-match accuracy of 0.15, indicating limited exact reproduction of the evaluated triples rather than widespread memorization of the KG.

B.10 Generated Causal Clues by DKCD

Example of Generated Causal Clues

Input Factors: Lung Cancer, Tuberculosis, Lung Disease, Chest X-ray, Dyspnea, Aisa Travel.

Generated Causal Clues:

1. **Tuberculosis → Lung Disease:** Tuberculosis infection can lead to severe pulmonary damage and respiratory disease.
2. **Lung Disease → Dyspnea:** Progressive lung disease is strongly associated with breathing difficulty and dyspnea.
3. **Lung Disease → Chest X-ray:** Patients with suspected lung disease commonly undergo chest X-ray examinations for clinical assessment.
4. **Tuberculosis → Asia Travel:** Recent travel to TB-endemic regions increases the likelihood of tuberculosis exposure.

Table 8: Sensitivity to random node and edge removal from the Diabetes KG using GPT-4o.

Removal ratio	Nodes	Edges	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
0%	22050	8643	$\mathbf{0.84 \pm 0.16}$	$\mathbf{0.71 \pm 0.08}$	$\mathbf{0.77 \pm 0.04}$	0.50 ± 0.17	$\mathbf{0.33 \pm 0.18}$	$\mathbf{0.40 \pm 0.18}$	$\mathbf{25.33 \pm 3.21}$
10%	19845	7779	0.66 ± 0.12	0.64 ± 0.15	0.65 ± 0.12	0.55 ± 0.22	0.26 ± 0.09	0.34 ± 0.12	34.33 ± 5.77
20%	17640	6914	0.60 ± 0.15	0.38 ± 0.18	0.46 ± 0.17	0.45 ± 0.39	0.15 ± 0.16	0.22 ± 0.21	36.00 ± 3.46
30%	15435	6050	0.45 ± 0.05	0.29 ± 0.00	0.35 ± 0.02	$\mathbf{0.78 \pm 0.19}$	0.13 ± 0.04	0.22 ± 0.06	39.00 ± 3.46

Table 9: Sensitivity to synthetic noise injected into the Diabetes KG using GPT-4o.

Noise ratio	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
0%	$\mathbf{0.84 \pm 0.16}$	0.71 ± 0.08	$\mathbf{0.77 \pm 0.04}$	0.50 ± 0.17	0.33 ± 0.18	0.40 ± 0.18	$\mathbf{25.33 \pm 3.21}$
10%	0.67 ± 0.04	$\mathbf{0.76 \pm 0.09}$	0.71 ± 0.05	0.59 ± 0.03	$\mathbf{0.34 \pm 0.05}$	$\mathbf{0.44 \pm 0.04}$	32.67 ± 2.31
20%	0.56 ± 0.05	0.21 ± 0.08	0.30 ± 0.08	0.47 ± 0.50	0.08 ± 0.08	0.12 ± 0.11	28.33 ± 2.89
30%	0.50 ± 0.00	0.21 ± 0.00	0.30 ± 0.00	$\mathbf{1.00 \pm 0.00}$	0.10 ± 0.04	0.19 ± 0.06	32.00 ± 0.00

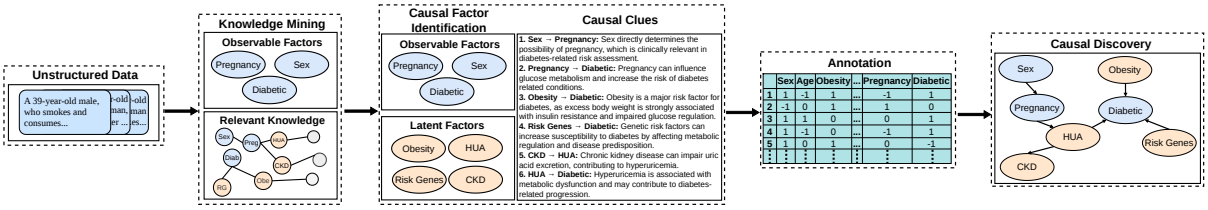


Figure 6: An example of the causal discovery process of DKCD on the Diabetes dataset.

Table 10: Full ablation study on DKCD components.

Dataset	Variant	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
Diabetes dataset	w/o CC	$\mathbf{0.84 \pm 0.16}$	$\mathbf{0.71 \pm 0.08}$	$\mathbf{0.77 \pm 0.04}$	0.49 ± 0.16	0.31 ± 0.13	0.38 ± 0.15	26.67 ± 2.52
	w/o LF	0.51 ± 0.17	0.24 ± 0.11	0.33 ± 0.14	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	34.00 ± 1.73
	w/o Both	0.43 ± 0.14	0.24 ± 0.11	0.31 ± 0.13	$\mathbf{0.67 \pm 0.58}$	0.05 ± 0.05	0.10 ± 0.09	35.67 ± 3.51
	DKCD	$\mathbf{0.84 \pm 0.16}$	$\mathbf{0.71 \pm 0.08}$	$\mathbf{0.77 \pm 0.04}$	0.50 ± 0.17	$\mathbf{0.33 \pm 0.18}$	$\mathbf{0.40 \pm 0.18}$	$\mathbf{25.33 \pm 3.21}$
Respiratory dataset	w/o CC	$\mathbf{0.74 \pm 0.13}$	$\mathbf{0.75 \pm 0.00}$	$\mathbf{0.74 \pm 0.07}$	0.73 ± 0.14	0.54 ± 0.08	0.62 ± 0.10	17.00 ± 2.65
	w/o LF	0.50 ± 0.17	0.42 ± 0.07	0.45 ± 0.11	$\mathbf{1.00 \pm 0.00}$	0.21 ± 0.14	0.32 ± 0.16	22.33 ± 4.51
	w/o Both	0.38 ± 0.16	0.30 ± 0.14	0.33 ± 0.15	0.33 ± 0.29	0.09 ± 0.08	0.14 ± 0.12	25.33 ± 1.53
	DKCD	$\mathbf{0.74 \pm 0.13}$	$\mathbf{0.75 \pm 0.00}$	$\mathbf{0.74 \pm 0.07}$	0.87 ± 0.23	$\mathbf{0.56 \pm 0.22}$	$\mathbf{0.68 \pm 0.21}$	$\mathbf{16.67 \pm 3.21}$

Table 11: Ablation study on design choices in DKCD.

Dataset	Variant	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
Diabetes dataset	w/o SM	0.36 ± 0.06	0.19 ± 0.04	0.23 ± 0.05	0.72 ± 0.49	0.08 ± 0.00	0.15 ± 0.01	34.67 ± 1.15
	w/o SP	0.60 ± 0.06	0.29 ± 0.00	0.39 ± 0.01	$\mathbf{0.78 \pm 0.39}$	0.18 ± 0.09	0.29 ± 0.14	29.67 ± 2.52
	DKCD	$\mathbf{0.84 \pm 0.16}$	$\mathbf{0.71 \pm 0.08}$	$\mathbf{0.77 \pm 0.04}$	0.50 ± 0.17	$\mathbf{0.33 \pm 0.18}$	$\mathbf{0.40 \pm 0.18}$	$\mathbf{25.33 \pm 3.21}$

Table 12: Impact of the top-r relevant knowledge contexts selected during semantic matching in the knowledge mining module on the performance of DKCD.

Dataset	r	NP↑	NR↑	NF↑	AP↑	AR↑	AF↑	ESHD↓
Diabetes dataset	5	0.33 ± 0.17	0.14 ± 0.07	0.20 ± 0.10	0.02 ± 0.03	0.03 ± 0.05	0.05 ± 0.08	36.67 ± 2.08
	10	0.55 ± 0.15	0.26 ± 0.09	0.36 ± 0.11	$\mathbf{0.83 \pm 0.29}$	0.13 ± 0.04	0.21 ± 0.06	32.00 ± 2.65
	20	$\mathbf{0.84 \pm 0.16}$	$\mathbf{0.71 \pm 0.08}$	$\mathbf{0.77 \pm 0.04}$	0.50 ± 0.17	$\mathbf{0.33 \pm 0.18}$	$\mathbf{0.40 \pm 0.18}$	$\mathbf{25.33 \pm 3.21}$
	50	0.61 ± 0.03	0.41 ± 0.15	0.48 ± 0.11	0.39 ± 0.24	0.16 ± 0.13	0.22 ± 0.17	33.00 ± 1.00

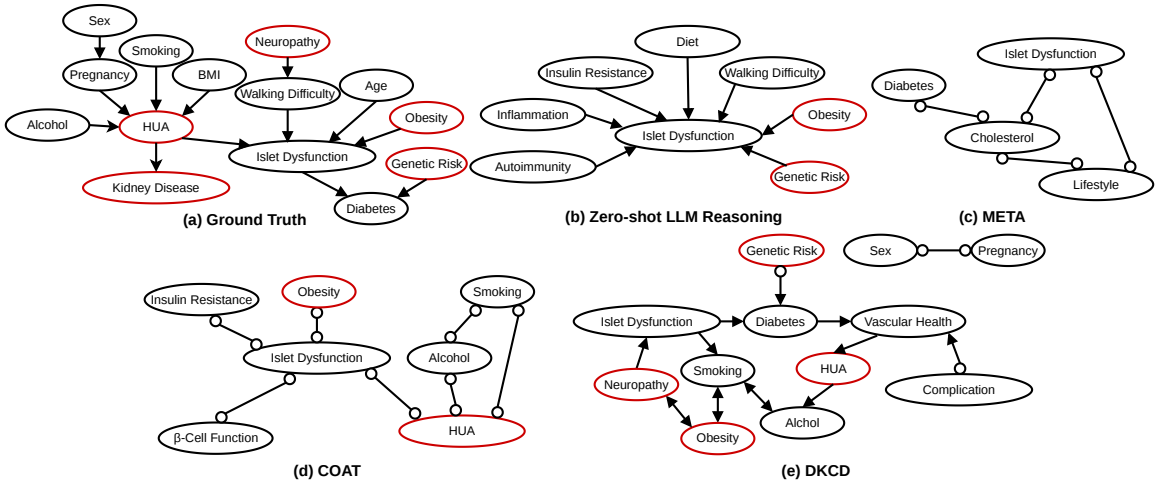


Figure 7: Causal graphs discovered with GPT-4o on the Diabetes dataset. Latent factors are marked in red.

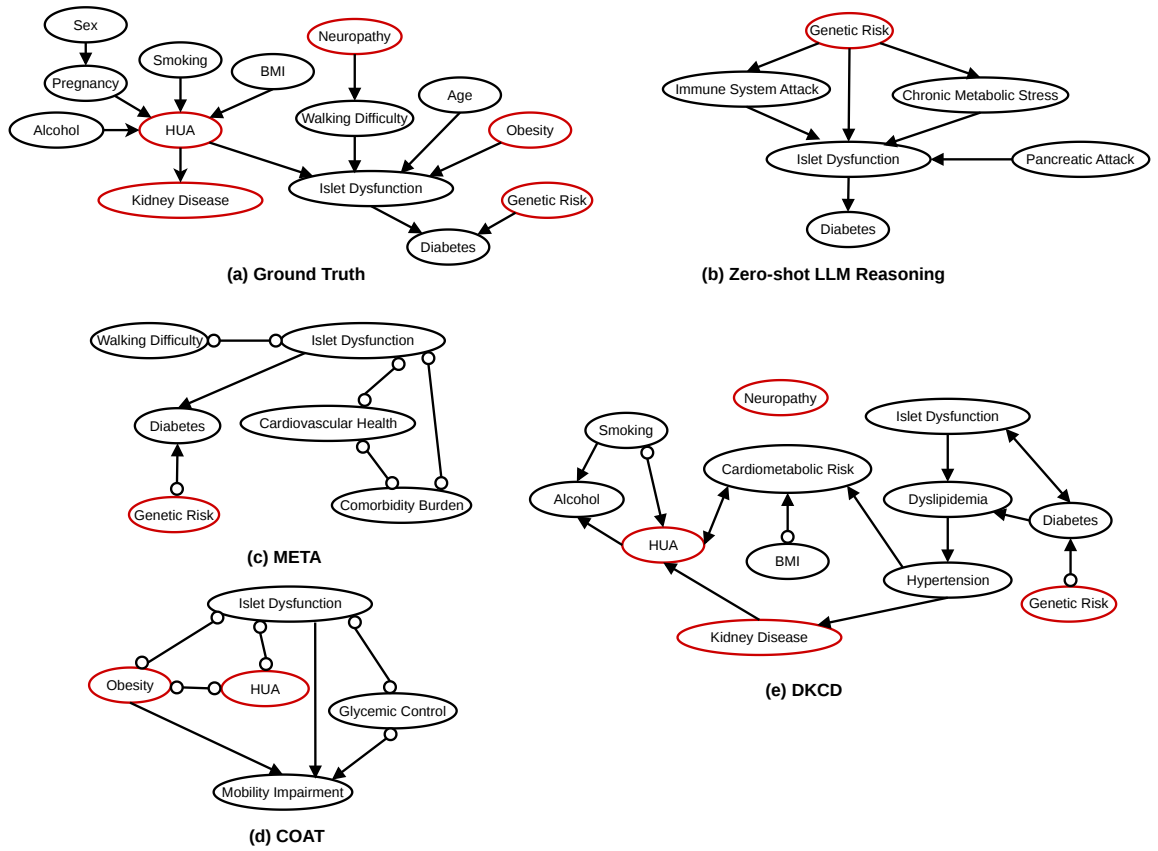


Figure 8: Causal graphs discovered with Gemini 2.5 Flash on the Diabetes dataset. Latent factors are marked in red.

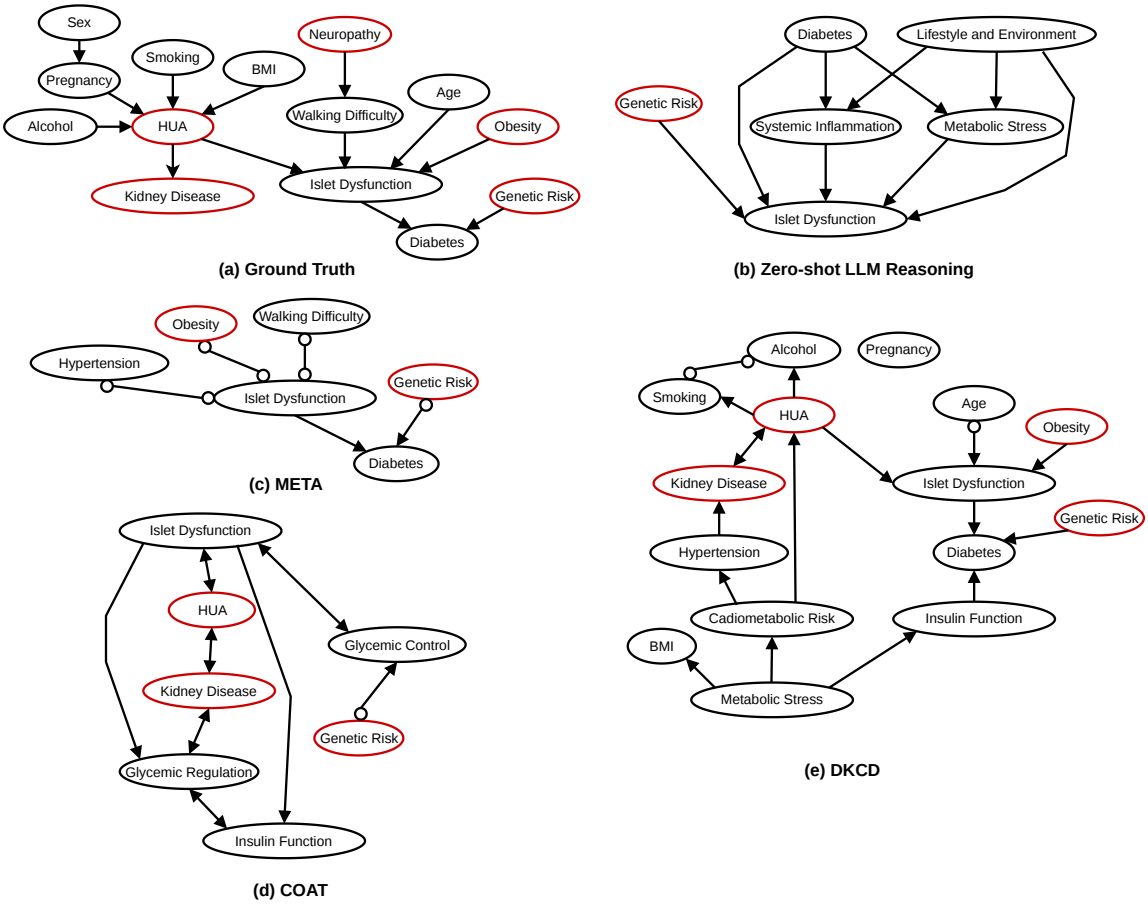


Figure 9: Causal graphs discovered with Grok-3 on the Diabetes dataset. Latent factors are marked in red.

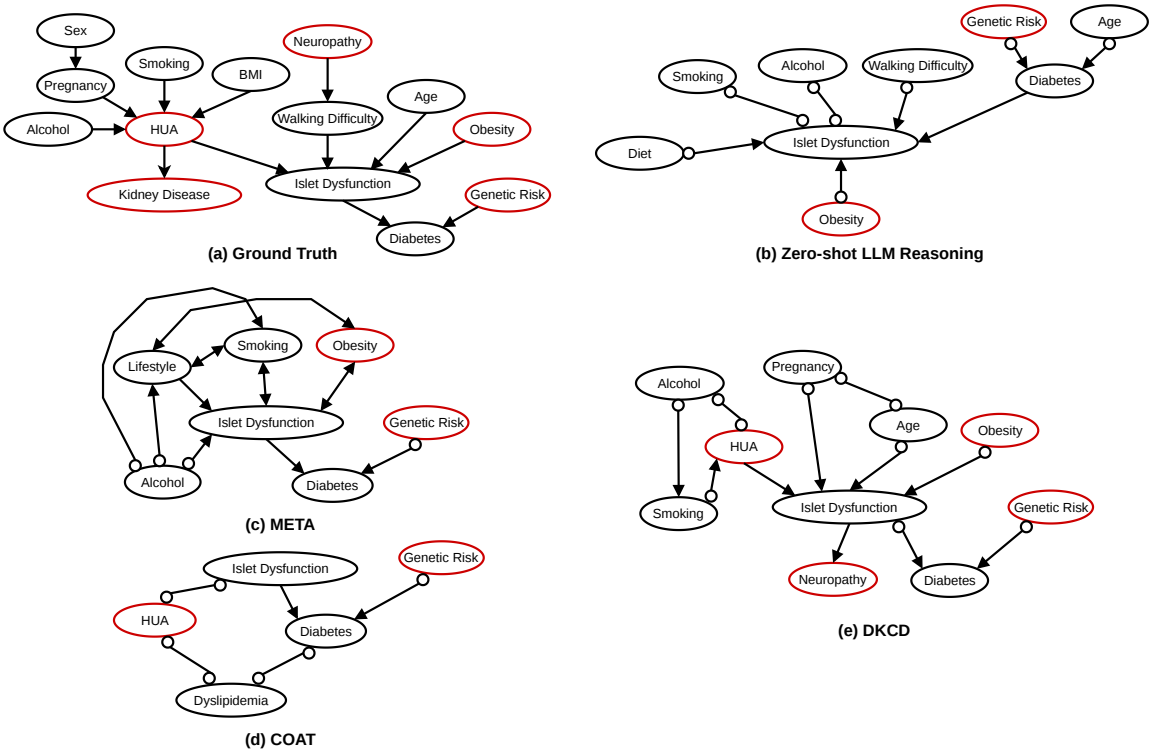


Figure 10: Causal graphs discovered with LLaMA 3-70B on the Diabetes dataset. Latent factors are marked in red.

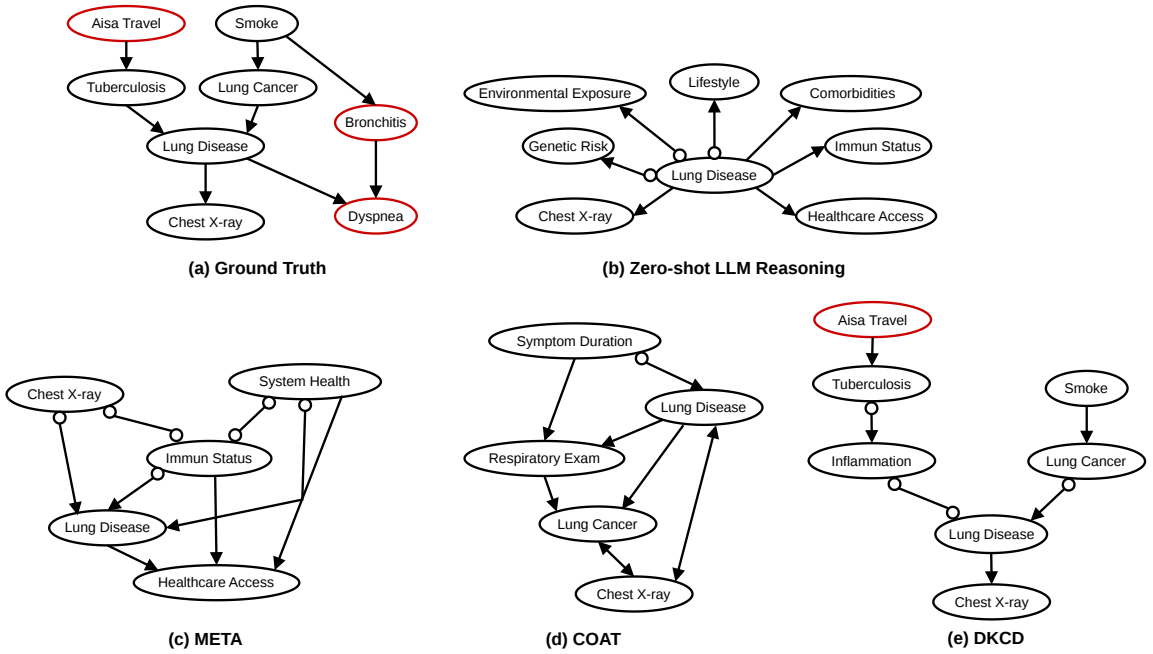


Figure 11: Causal graphs discovered with GPT-4o on the Respiratory dataset. Latent factors are marked in red.

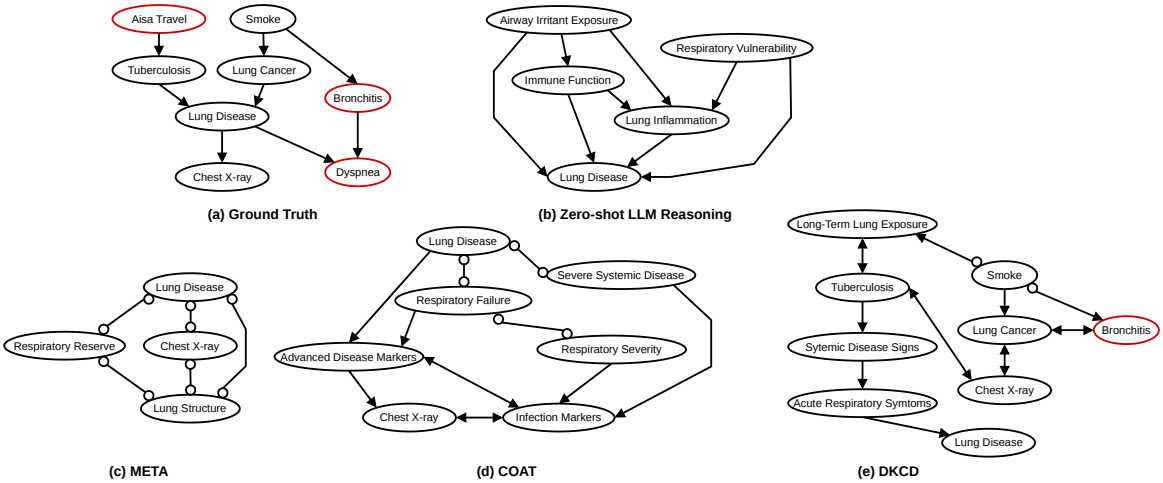


Figure 12: Causal graphs discovered with Gemini 2.5 Flash on the Respiratory dataset. Latent factors are marked in red.

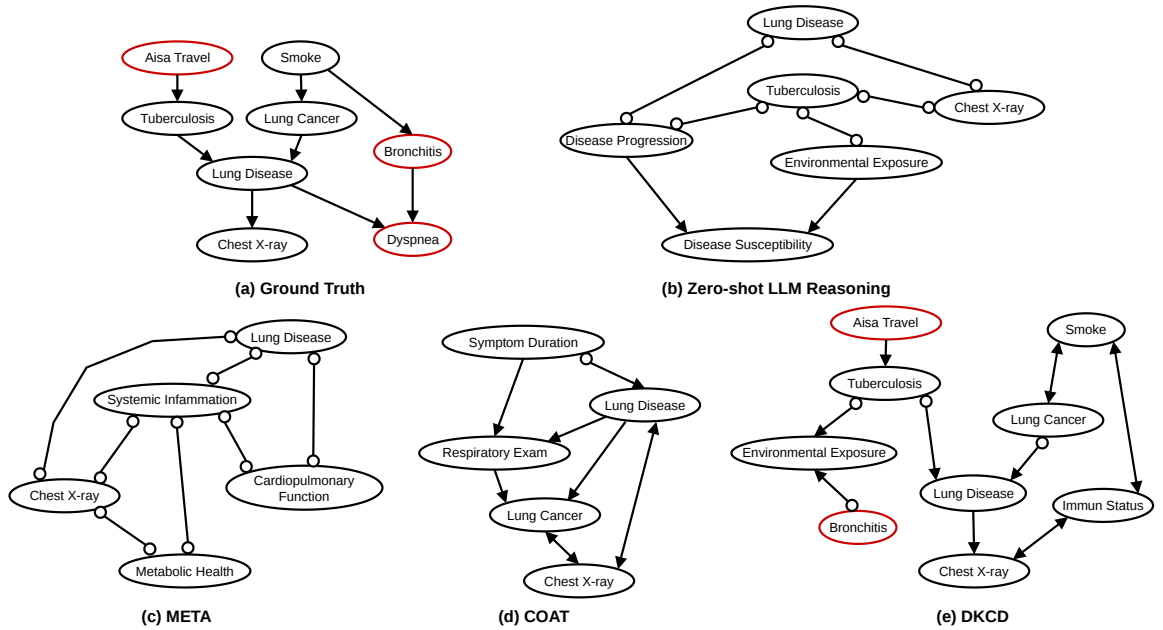


Figure 13: Causal graphs discovered with Grok-3 on the Respiratory dataset. Latent factors are marked in red.

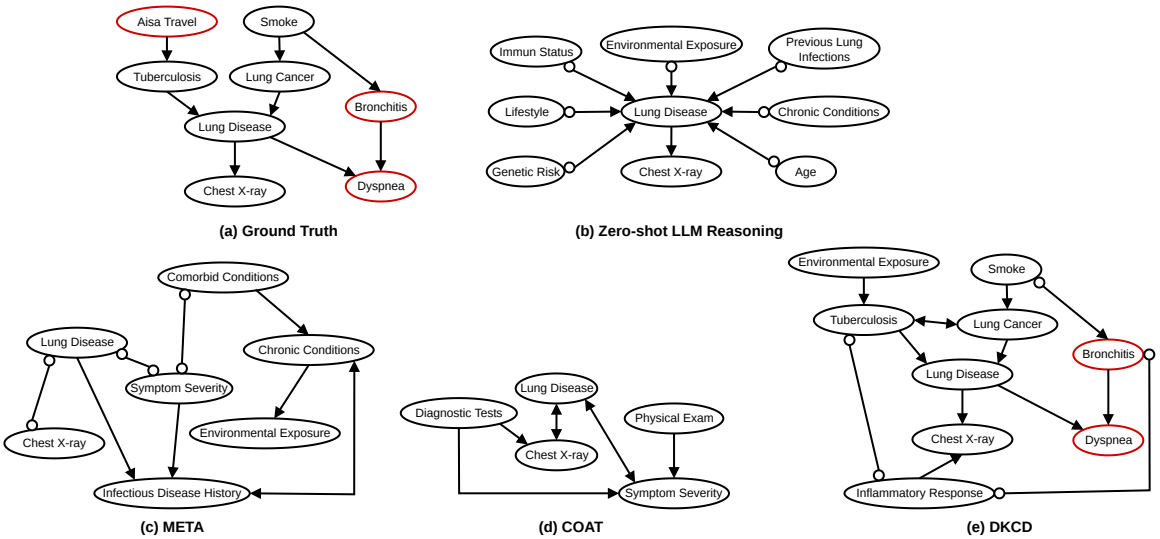


Figure 14: Causal graphs discovered with LLaMA 3-70B on the Respiratory dataset. Latent factors are marked in red.