

ods generate reasoning structures only at inference time through prompting, without any training signal to correct mis-specified causal relationships. Consequently, when an LLM incorrectly identifies A as causing B (when B actually causes A), or fails to recognize a confounding variable C that influences both, no gradient flows back to fix these fundamental errors (Wang et al. 2023; Yao et al. 2023; Besta et al. 2024). As a result, they cannot block spurious backdoor paths or guarantee counterfactual consistency, leaving LLMs still vulnerable to logically inconsistent hallucinations (Wang et al. 2023; Yao et al. 2023; Besta et al. 2024). The mathematical constraints further compound this problem. Causal relationships inherently form a DAG that encodes multiple interconnected variables with conditional dependencies and multiple pathways of influence. A linear chain or even a tree structure cannot adequately represent scenarios where a variable influences multiple outcomes simultaneously or where effects depend on the interaction of multiple causes, both fundamental characteristics of causal DAG. This structural mismatch means that prompt-only variants such as CoT, ToT, GoT, and DoT cannot, by design, supervise LLMs to learn causal edge semantics, limiting their ability to enforce conditional independencies required for true causal inference.

To address this gap, we propose **Causal-DAG Construction and Reasoning** (CDCR-SFT), a supervised fine-tuning framework that trains LLMs to first construct a variable-level causal DAG and then reason over that graph. The training of CDCR-SFT requires data with a causal DAG as well as the corresponding reasoning on top of that. Therefore, we introduce CausalDR (Causal-DAG and Reasoning), the first dataset specifically designed to train LLMs in simultaneous causal DAG construction and graph-based reasoning. Building upon the CLADDER dataset (Jin et al. 2023), which provides causal questions with a causal DAG, we develop an automated generation and validation pipeline using DeepSeek-R1 (DeepSeek-AI 2025). This pipeline ensures high-quality data generation through question-answer consistency checks. Each sample in CausalDR comprises (1) an input question, (2) a causal DAG that explicitly describes variables and their relationships, (3) a graph-based reasoning trace that navigates the causal structure, and (4) the final answer. As shown in Fig. 1, our experiments demonstrate that CDCR-SFT can address our research question by both improving causal reasoning capabilities and mitigating the logically inconsistent hallucinations across multiple benchmarks. This indicates that, rather than solely pursuing larger model sizes or more training data or longer cot, we can achieve more trustworthy LLMs by equipping them with structured reasoning capabilities that align with the underlying causal nature of real-world problems.

The major contributions of this work are three-fold. First, we introduce CDCR-SFT, a supervised fine-tuning framework that shifts how LLMs approach causal reasoning by moving from sequential CoT to DAG-based inference. It trains models to construct a causal DAG that properly encodes both causal directionality and conditional independence relationships, enabling them to perform structured

reasoning over these graphs rather than being constrained by linear reasoning paths. Second, we present CausalDR, a dataset containing 25,368 high-quality samples for teaching LLMs to generate causal DAG construction and reason on top of the DAG. Third, we demonstrate that explicit causal structure modeling can not only improve causal reasoning but also mitigate hallucinations in LLMs.

2 Related Works

Reasoning and Causal Limitations in LLMs LLMs employ structured reasoning methods such as Chain-of-Thought (CoT) prompting, which generates intermediate steps alongside final answers (Wei et al. 2022); Self-Consistency (CoT-SC), which samples multiple reasoning chains for robustness; Tree-of-Thoughts (ToT), which branches into alternative solution paths (Yao et al. 2023); and Graph-of-Thoughts (GoT), which links subproblems as nodes in a simple graph (Besta et al. 2024). However, these methods treat inference as linear sequences or trees and cannot represent directed acyclic graph (DAG) needed for causal analysis, where edges denote cause–effect relations and support interventions and counterfactual reasoning. Benchmarks such as CausalBench show that LLMs struggle with intervention and counterfactual queries, failing to predict outcomes of hypothetical changes (Wang 2024), and synthetic tests confirm that models rely on surface text patterns rather than true cause–effect relations (Ma 2024).

Hallucination Reduction and Causal Supervised Fine-Tuning Complex reasoning tasks can exacerbate hallucinations in LLMs, as models often rely on surface-level correlations rather than true causal structure (Bagheri et al. 2024). Traditional mitigation—external knowledge checks or post-hoc filters—only corrects errors after generation and does not strengthen the model’s internal inference process (Wang 2024). Recent studies have demonstrated that task-specific fine-tuning significantly improves LLM performance on specialized benchmarks (Han et al. 2024; Liu et al. 2025). In particular, supervised fine-tuning (SFT) with low-rank adapters (LoRA) (Hu et al. 2022) reshapes internal reasoning by training models on structured targets. In this study, we extend this paradigm by using the CausalDR dataset’s annotated DAG and stepwise reasoning to teach the model to first construct a causal graph and then perform graph-based inference, thereby reducing hallucinations and improving consistency.

3 Methods

3.1 CDCR-SFT

CDCR-SFT is a supervised fine-tuning framework that trains LLMs to explicitly perform causal reasoning through causal-DAG construction and reasoning. Specifically, LLMs learn to construct a causal DAG by identifying causal variables from input queries, then perform structured reasoning over the DAG, and finally generate answers, as shown in Fig. 2.

Existing structured reasoning methods, such as CoT, ToT, and GoT, generally produce reasoning paths at the linguis-

tic level, while CDCR-SFT performs reasoning at the causal level. The proposed framework consists of two main components: causal reasoning and causal construction. Causal reasoning involves generating causal DAGs based on input queries, while causal construction involves performing structured reasoning over the generated DAGs. The proposed framework is designed to be more efficient and accurate than existing methods, as it leverages the strengths of both causal reasoning and causal construction.

Bu boşluğu gidermek için, Büyük Dil Modellerini (BDM) öncelikle değişken düzeyinde bir nedensel Yönlendirilmiş Döngüsel Olmayan Çizelge (YADÇ) oluşturmak ve ardından bu çizelge üzerinde akıl yürütme eğiten denetimli bir ince ayar çerçevesi olan Nedensel YADÇ Oluşturma ve Akıl Yürütme (CDCR-SFT) sistemini öneriyoruz. CDCR-SFT’nin eğitimi, bir nedensel YADÇ içeren verilerin yanı sıra bunun üzerinde ilgili akıl yürütme gereklidir. Bu nedenle, büyük dil modellerini eş zamanlı nedensel YADÇ oluşturma ve grafik tabanlı akıl yürütme konusunda eğitmek için özel olarak tasarlanmış ilk veri kümeleri olan CausalDR’yi (Nedensel Yönlendirilmiş Döngüsel Olmayan Çizelge ve Akıl Yürütme) sunuyoruz. Nedensel soruları bir nedensel Yönlendirilmiş Döngüsel Olmayan Çizelge ile sunan CLADDER veri kümelerini (Jin vd. 2023) temel olarak, DeepSeek-R1 (DeepSeek-AI 2025) kullanarak otomatik bir üretim ve doğrulama süreci geliştiriyoruz. Bu süreç, soru-cevap tutarlılığı kontrolleri aracılığıyla yüksek kaliteli veri üretimi garanti eder. CausalDR’deki her örnek (1) bir giriş sorusu, (2) değişkenleri ve ilişkilerini açıkça tanımlayan bir nedensel Yönlendirilmiş Döngüsel Olmayan Çizelge, (3) nedensel yapıyı yönlendiren grafik tabanlı bir akıl yürütme izi ve (4) nihai cevaptan oluşur. Şekil 1’de gösterildiği gibi, deneylerimiz CDCR-SFT’nin hem nedensel akıl yürütme yeteneklerini geliştirmek hem de birden fazla kıyaslamada mantıksal olarak tutarsız halüsinasyonları azaltarak araştırma sorumuzu ele alabileceğini ortaya koymaktadır. Bu durum, yalnızca daha büyük model boyutları, daha fazla eğitim verisi veya daha uzun Düşünce Zinciri arayışına girmek yerine, büyük dil modellerini gerçek dünya problemlerinin temelindeki nedensel doğayla uyumlu yapılandırılmış akıl yürütme yetenekleriyle donatarak daha güvenilir hale getirebileceğimizi göstermektedir.

Bu çalışmanın başlıca katkıları üç yönlüdür. İlk olarak, Büyük Dil Modellerinin (BDM) nedensel akıl yürütmeye yaklaşımını sıralı Düşünce Zinciri’nden (CoT) Yönlendirilmiş Döngüsel Çizge (DAG) tabanlı çıkarma kaydırır, denetimli bir ince ayar çerçevesi olan CDCR-SFT’yi sunuyoruz. Bu çerçeve, modelleri hem nedensel yönlüğü hem de koşullu bağımsızlık ilişkilerini doğru bir şekilde kodlayan bir nedensel YADÇ oluşturmaları için eğiterek, yapılandırılmış

doğrusal akıl yürütme yollarıyla sınırlı kalmak yerine bu çizelgeler üzerinde akıl yürütme yapmalarını sağlar. İkinci olarak, BDM’lere nedensel YADÇ inşası üretmeyi ve YADÇ üzerinde akıl yürütmemeyi öğretmek için 25.3 68 yüksek kaliteli örnek içeren bir veri kümeleri olan CausalDR’yi tanıtıyoruz. Üçüncü olarak, açık nedensel yapı modellemesinin yalnızca nedensel akıl yürütmemeyi iyileştirmekle kalmayıp aynı zamanda BDM’lerdeki halüsinasyonları da azaltabileceğini gösteriyoruz.

2 İlgili Çalışmalar

BDM’erde Akıl Yürütme ve Nedensel Kısıtlamalar BDM’ler, nihai cevapların yanı sıra adımlar üreten Düşünce Zinciri (CoT) yönlendirmesi gibi yapılandırılmış akıl yürütme yöntemleri kullanır (Wei vd. 2022); Sağlamlık için birden çok akıl yürütme zincirini örnekleyen Kendi Kendine Tutarlılık (Düşünce Zinciri-SC); Alternatif çözüm yollarına ayrılan Düşünce Ağacı (ToT) (Yao vd. 2023);

ve alt problemleri basit bir çizgedeki düğümler olarak bağlayan Düşünce Çizgesi (GoT) (Besta ve diğerleri. 2024). Ancak, bu yöntemler karışımı doğrusal diziler veya ağaçlar olarak ele alır ve nedensel analiz için gerekli, kenarların neden-sonuç ilişkilerini gösterdiği ve müdahaleleri ile karşılıklı akıl yürütmemeyi desteklediği bir yönlendirilmiş döngüsel çizgeyi (DAG) temsil edemezler. CausalBench gibi kıyaslama testleri, BDM’lerin müdahale ve karşılıklı sorgularla zorlandığını, varsayımsal değişikliklerin sonuçlarını tahmin edemediğini göstermektedir (Wang 2024) ve sentetik testler, modellerin gerçek neden-sonuç ilişkileri yerine yüzeysel metin kalıplarına dayandığını doğrulamaktadır (Ma 2024).

Halüsinasyon Azaltma ve Nedensel Denetimli İnce Ayar Karmaşık akıl yürütme görevleri, modellerin genellikle gerçek nedensel yapıdan ziyade yüzeysel korelasyonlara dayanması nedeniyle BDM’lerdeki halüsinasyonları şiddetlendirilebilir (Bagheri ve diğerleri. 2024). Geleneksel azaltma yöntemleri – harici bilgi kontrolleri veya sonradan uygulanan filtreler – hataları yalnızca üretim sonrasında düzeltir ve modelin dahili çıkarım sürecini güçlendirmez (Wang 2024). Son çalışmalar, görevde özel ince ayarın, BDM’lerin uzmanlaşmış karşılaştırma testlerindeki performansını önemli ölçüde artırdığını göstermiştir (Han ve diğerleri 2024; Liu ve diğerleri 2025). Özellikle, düşük dereceli adaptörlerle (LoRA) (Hu ve diğerleri 2022) denetimli ince ayar (SFT), modelleri yapılandırılmış hedefler üzerinde eğitemek dahili akıl yürütme yeniden şekillendirir. Bu çalışmada, CausalDR veri kümelerinin açıklanmış nedensel Yönlendirilmiş Döngüsel Olmayan Çizelge’sini ve adım adım akıl yürütmesini kullanarak modeli öncelikle bir nedensel çizge oluşturmaya ve ardından çizge tabanlı çıkarım yapmaya öğretmeyi, böylece halüsinasyonları azaltmayı ve tutarlılığı artırmayı amaçlayarak bu paradigmayı genişletiyoruz.

3 Yöntemler

3.1 CDCR-SFT

CDCR-SFT, Büyük Dil Modellerini (BDM) nedensel Yönlendirilmiş Döngüsel Çizge oluşturma ve akıl yürütme yoluyla açıkça nedensel akıl yürütme gerçekleştirmeleri için eğiten denetimli bir ince ayar çerçevesidir. Özellikle, BDM’ler, Şekil 2’de gösterildiği gibi, girdi sorgularından nedensel değişkenleri tanımlayarak bir nedensel Yönlendirilmiş Döngüsel Olmayan Çizelge (YADÇ) oluşturmayı, ardından YADÇ üzerinde yapılandırılmış akıl yürütme gerçekleştirmeyi ve son olarak yanıtlar üretmeyi öğrenir. Düşünce Zinciri (CoT), Düşünce Ağacı (ToT) ve Oyun Teorisi (GoT) gibi mevcut yapılandırılmış akıl yürütme yöntemleri genellikle dilsel düzeyde akıl yürütme yolları üretir.

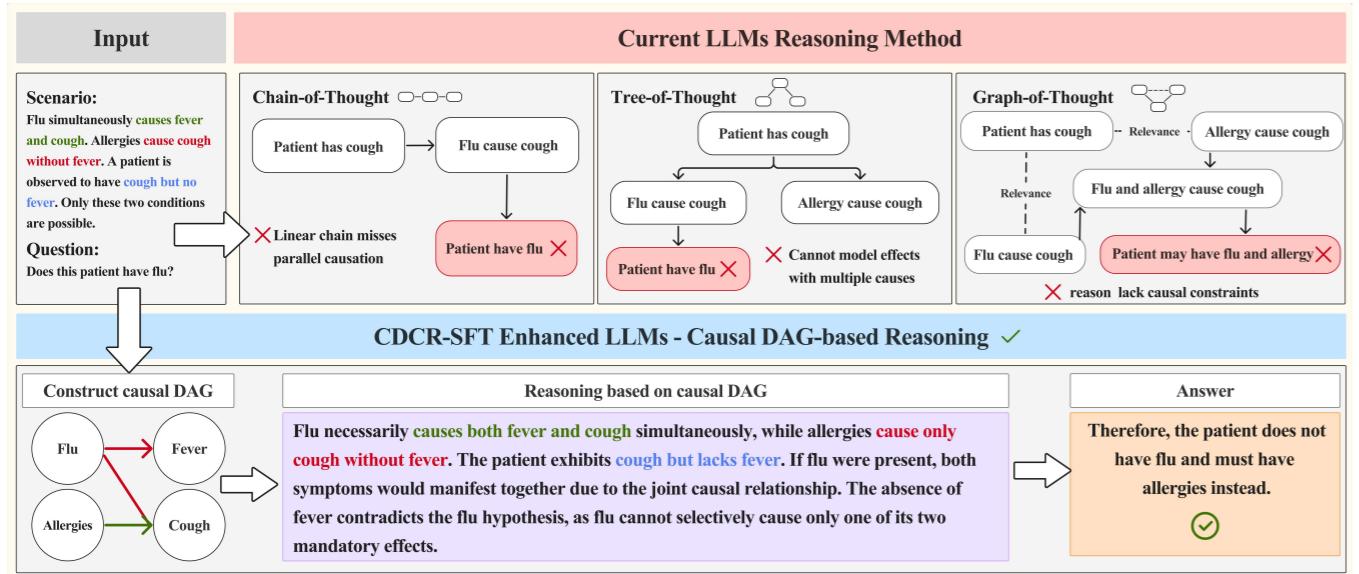


Figure 2: Comparison of reasoning approaches: Existing methods (CoT, ToT, GoT) operate at linguistic/semantic levels without explicit causal structure; Our CDCR-SFT constructs a variable-level causal DAG with directed edges representing causal relationships, enabling principled causal inference through graph-based reasoning.

tic token or semantic levels without modeling the underlying causal structures among variables. Table 1 provides a detailed comparison of key capabilities between our proposed CDCR-SFT framework and existing reasoning methods. Mathematically, CoT generates reasoning paths as linear reasoning sequences $S_{\text{CoT}} = (p_1, \dots, p_n, y)$, ToT forms branching reasoning trees $S_{\text{ToT}} = \text{Tree}(p_1, \dots, p_n, y)$, and GoT creates semantic-level reasoning graphs $S_{\text{GoT}} = \text{Graph}(p_1, \dots, p_n, y)$. CDCR-SFT outputs a DAG-based

Aspect	CDCR-SFT (ours)	CoT	ToT	GoT
Reasoning aligned with causal relationships	✓	✗	✗	✗
Explicit causal training signal	✓	✗	✗	✗
Supports multi-parent causes	✓	✗	✗	✗ (no causal)
Captures conditional independencies	✓	✗	✗	✗
Captures interventions	✓	✗	✗	✗
Captures counterfactuals	✓	✗	✗	✗
Effective hallucination mitigation	✓	✗	✗	✗

Table 1: Comparison of key capabilities between CDCR-SFT and existing reasoning methods.

reasoning process $S_{\text{CDCR-SFT}} = (G, P, y)$, where $G = (V, E)$ denotes the causal DAG encoding causal directionality and conditional independence relationships, $P = (p_1(G), \dots, p_n(G))$ represents reasoning steps that adhere strictly to causal structures in G , and y is the final inferred answer. Specifically, in the textual encoding of the causal DAG G , each causal variable is clearly represented as a node described in natural language, including detailed descriptions of the primary events. The causal relationships among these variables are encoded as directed edges, explicitly indicating directional influences. An illustrative example of textual DAG encoding is provided in Fig. 3.

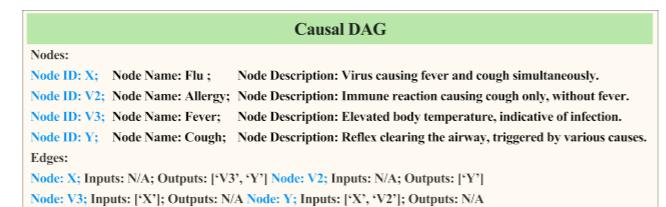
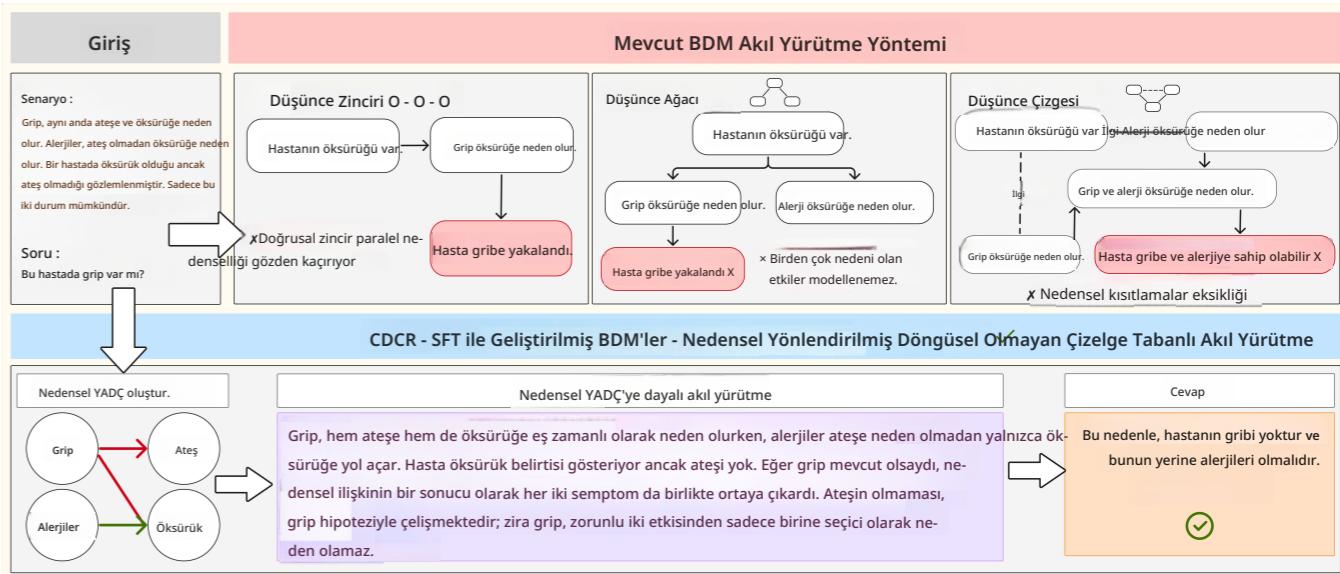


Figure 3: Textual representation of the causal DAG in Fig. 2.

3.2 Dataset Construction

Causal-DAG and Reasoning (CausalDR) Dataset To train LLMs in simultaneous causal DAG construction and graph-based reasoning, we require datasets explicitly providing supervision for both. However, existing causal datasets (Gordon, Kozareva, and Roemmele 2012; Tandon et al. 2019; Du et al. 2022) either omit explicit causal relationships altogether or, as exemplified by CLADDER (Jin et al. 2023), offer mathematically rigorous yet semantically sparse causal graphs and algebraic formulations, lacking clear natural-language reasoning paths linking structures to answers (A CLADDER example is provided in Appx. A.1).

We introduce **CausalDR**, the first large-scale annotated dataset explicitly designed for supervised fine-tuning of LLMs in simultaneous causal DAG construction and structured causal reasoning. Each training sample in CausalDR consists of: (1) an instruction specifying the task, (2) an input question or scenario, and (3) a coherent output comprising three components: a text-based causal DAG G , a reasoning path $P == (p_1(G), \dots, p_n(G))$ based on G , and a final answer y derived through structured inference (A detailed example see Appx. A.2).



Şekil 2: Akıl yürütme yaklaşımlarının karşılaştırılması: Mevcut yöntemler (CoT, ToT, GoT) açık nedensel yapı olmaksızın dilbilimsel/antılamalı düzeylerde çalışır; CDCR-SFT'mız, nedensel ilişkileri temsil eden yönlendirilmiş kenarlarla değişken düzeyinde bir nedensel YADÇ oluşturur ve grafik tabanlı akıl yürütme yoluyla ilkel nedensel çıkarım sağlar.

değişkenler arasındaki temel nedensel yapıları modellemeden anlamsal düzeylerde veya dilsel düzeylerde. Tablo 1, önerilen CDCR-SFT çerçevesimiz ile mevcut akıl yürütme yöntemleri arasındaki temel yeteneklerin ayrıntılı bir karşılaştırmasını sunmaktadır. Matematiksel olarak, CoT akıl yürütme yollarını doğrusal akıl yürütme dizileri olarak üretir $S_{\text{CoT}} = (p_1, \dots, p_n, y)$, ToT dallanmış akıl yürütme ağaçları oluşturur $S_{\text{ToT}} = \text{Tree}(p_1, \dots, p_n, y)$ ve GoT anlamsal düzeye de akıl yürütme grafikleri oluşturur $S_{\text{GoT}} = \text{Graph}(p_1, \dots, p_n, y)$. CDCR-SFT DAG tabanlı bir çıktı üretir

Düzenli Yönlendirilmiş Dönüşel Oluşan Çizge	
Düzenli YADÇ'ye dayalı akıl yürütme	
Düğüm Kimliği: X; Düğüm Adı: Grip;	Düğüm Açıklaması: Aynı anda ateş ve öksürüğe neden olan virüs.
Düğüm Kimliği: V2; Düğüm Adı: Alerji;	Düğüm Açıklaması: Sadece öksürüğe neden olan, ateşle yol açmayan bağıksızlık tekipi.
Düğüm Kimliği: V3; Düğüm Adı: Ateş;	Düğüm Açıklaması: Enfeksiyonu gösteren yüksek vücut sıcaklığı.
Düğüm Kimliği: Y; Düğüm Adı: Öksürük;	Düğüm Açıklaması: Çeşitli nedenlerle tetiklenen, havayolu temizleyen refleks.
Kenarlar:	
Düğüm: X; Girdiler: N/A; Çıktılar: ['V3', 'Y'] Düğüm: V2; Girdiler: N/A; Çıktılar: ['Y']	
Düğüm: V3; Girdiler: ['X']; Çıktılar: N/A Düğüm: Y; Girdiler: ['X', 'V2']; Çıktılar: N/A	

Şekil 3: Şekil 2'deki nedensel YADÇ'nın metinsel gösterimi.

Yön	CDCR-SFT (bizim) Düşünce Zinciri	Oyun Teorisı
Nedensel ilişkilerle uyumlu akıl yürütme	✓	✗
Açık nedensel eğitim sinyali	✓	✗
Çoklu ebeveyn nedenlerini destekler	✓	✗ (nedensel değil)
Kosullu bağımsızlıklar yakalar	✓	✗
Müdahaleleri yakalar	✓	✗
Karşılıklı durumları yakalar	✓	✗
Etkili halusinasyon azaltma	✓	✗

Tablo 1: CDCR-SFT ve mevcut akıl yürütme yöntemleri arasındaki temel yeteneklerin karşılaştırılması.

akıl yürütme süreci $S_{\text{CDCR-SFT}} = (G, P, y)$, burada $G = (V, E)$ nedensel yönlendirmeyi ve koşullu bağımsızlık ilişkilerini kodlayan nedensel YADÇ'yi belirtir, $P = (p_1(G), \dots, p_n(G))$ G 'deki nedensel yapılara sıkı sıkıya bağlı kalan akıl yürütme adımlarını temsil eder ve y sonuç olarak çıkarılan cevaptır. Özellikle, nedensel YADÇ G 'nin metinsel kodlamasında, her nedensel değişken, ana olayların ayrıntılı açıklamaları da dahil olmak üzere doğal dilde tanımlanan bir düğüm olarak açıkça temsil edilir. Bu değişkenler arasındaki nedensel ilişkiler, yönlü kenarlar olarak kodlanır ve yönlü etkileri açıkça gösterir. Metinsel Yönlendirilmiş Dönüşel Çizge (DAG) kodlamasına açıklayıcı bir örnek Şekil 3'te sunulmuştur

CausalDR'yi tanıtıyoruz; bu, Büyük Dil Modellerinin (BDM'lerin) eş zamanlı nedensel Yönlendirilmiş Dönüşel Çizge (YADÇ) inşası ve yapılandırılmış nedensel akıl yürütmesi için özel olarak tasarlanmış ilk büyük ölçekli açıklamalı veri kümeleridir. CausalDR'deki her eğitim örneği şunları içerir: (1) görevi belirten bir talimat, (2) bir giriş sorusu veya senaryosu ve (3) üç bileşenden oluşan tutarlı bir çıktı: metin tabanlı bir nedensel YADÇ G , G 'ye dayalı bir akıl yürütme yolu $P == (p_1(G), \dots, p_n(G))$ ve yapılandırılmış çıkarım yoluyla türetilmiş nihai bir cevap y (Ayrıntılı bir örnek için Ek A.2'ye bakınız).

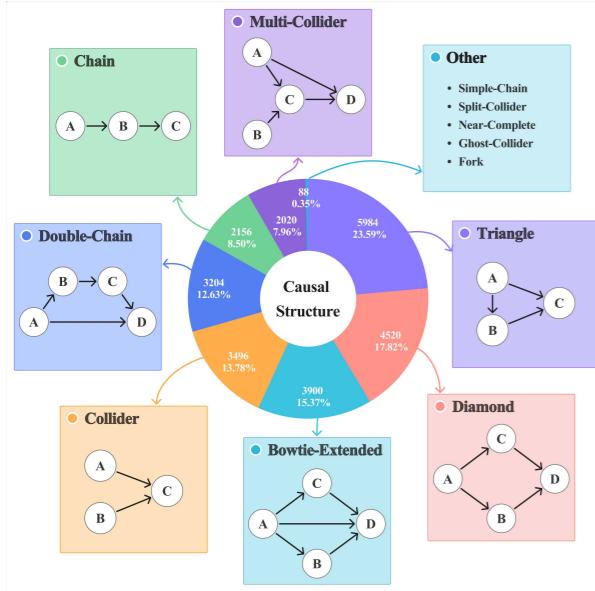


Figure 4: Proportional Distribution of 12 Canonical Causal DAG Structures in the CausalDR Dataset.

We construct CausalDR based on the CLADDER dataset (Jin et al. 2023), partitioning it into training and test sets based on unique identifiers (`graph_id` and `story_id`) to prevent information leakage. And then using the DeepSeek-R1 (DeepSeek-AI 2025) (temperature=0.6, max tokens=8192, details in Appx. A.3), we developed an automated pipeline (pseudocode provided in Appx. A.5) to generate and validate high-quality training samples for the CausalDR dataset. Specifically, we designed a prompt (details in Appx. A.4) that contains a mathematically accurate causal DAG expressed in formal notation, instructing DeepSeek-R1 to produce JSON-formatted outputs for each CLADDER sample. Each output explicitly: (1) causal nodes with clear semantic descriptions, and causal edges specifying incoming and outgoing relationships, (2) a step-by-step reasoning path that explicitly references the constructed causal DAG, and (3) the final inferred answer. To ensure quality, we implemented a validation mechanism comparing model-generated answers against the original ground-truth answers provided by CLADDER. If a generated answer did not match the ground-truth after multiple validation attempts, the sample was manually reviewed or discarded. Through this process, we obtained a high-quality dataset of 6,357 validated samples.

To further enhance dataset diversity and generalization, we introduced a Causal DAG Augmentation technique. Specifically, given an original causal DAG $G = (V, E)$, we randomly permuted the order of causal nodes and edges using permutation functions $\pi_v(\cdot)$ and $\pi_e(\cdot)$, respectively, to create diverse augmented variants: $V_{\text{aug}} = \pi_v(V)$, $E_{\text{aug}} = \pi_e(E)$, $G_{\text{aug}} = (V_{\text{aug}}, E_{\text{aug}})$. We applied this permutation procedure four times per original DAG G , each time pairing the permuted DAG G_{aug} with the original reasoning path P and answer y . This expanded the initial dataset from 6,357

samples to 25,368 augmented training examples.

Fig. 4 shows the proportional distribution of the 12 canonical causal DAG structures within the CausalDR dataset. These structures cover diverse causal configurations, including simple Chains (e.g., Chain, Double-Chain), Confounding structures (e.g., Triangle, Fork), Colliders (e.g., Collider, Multi-Collider), and more intricate multi-path interactions (e.g., Diamond, Bowtie-Extended). The diverse representation of these key causal mechanisms enables effective generalization of causal reasoning capabilities in large language models.

Auxiliary Instruction following Data To prevent the model from over-focusing on the causal task and degrading the linguistic generalization ability, we randomly select 10,000 Alpaca (Taori et al. 2023) examples and mix them with the CausalDR dataset during supervised fine-tuning to ensure the overall linguistic ability and generalization performance of the model.

3.3 Supervised Fine-tuning Procedure

During supervised fine-tuning, LLM learns to generate the structured causal DAG inference sequence $S_{\text{CDCR-SFT}} = (G, P, y)$. The optimization objective is formulated as a negative log-likelihood loss: $\mathcal{L}_{\text{CDCR-SFT}} = -\sum_{t=1}^{|S|} \log P(s_t | s_{<t}, X)$, where s_t denotes the t -th token in the ground-truth sequence S , and $s_{<t}$ represents all tokens before position t .

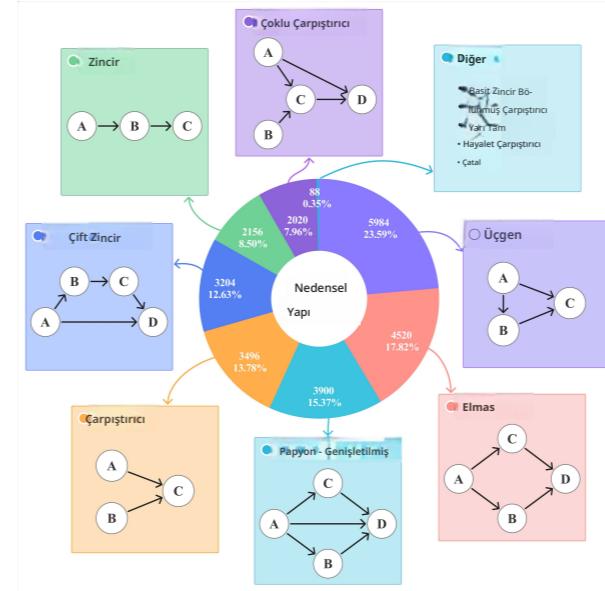
Critically, whenever the model-generated sequences deviate from the ground-truth causal DAG structure—such as introducing reversed causal edges, omitting essential causal variables, or adding extraneous causal relationships—explicit gradient signals immediately correct these inaccuracies. This supervision ensures that the model internalizes correct causal directionality, conditional independence properties, and intervention semantics required for accurate causal reasoning.

For computational efficiency, we applied Low-Rank Adaptation (LoRA) (Hu et al. 2022) during fine-tuning, updating only a small number of low-rank parameters inserted into each layer, while freezing the original pretrained LLM parameters. Through this fine-tuning procedure, CDCR-SFT trains the model to construct an accurate causal DAG and perform structured reasoning explicitly constrained by the causal relationships defined in these graphs, thereby improving the logical consistency of the LLM outputs and mitigating hallucinations.

4 Experiments

4.1 Experimental Setup

Base LLMs and Reasoning Methods We select four pretrained LLMs for evaluation: (1) Llama-3.1-8B-Instruct (Grattafiori et al. 2024), (2) DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI 2025), (3) Baichuan2-7B-Chat (Baichuan 2023), and (4) Mistral-7B-Instruct-v0.2 (Jiang et al. 2023). Additionally, we compare our CDCR-SFT method against five baseline reasoning approaches: Zero-shot-CoT(CoT)(Kojima et al. 2023), Chain-of-Thought Self-Consistency (CoT-SC)(Wang et al.



Şekil 4: CausalDR Veri Kümesindeki 12 Kanonik Nedensel Yönlendirilmiş Döngüsel Olmayan Çizge Yapılarının Oransal Dağılımı.

CausalDR'ı, CLADDER veri kümesine (Jin vd. 2023) dayanarak oluşturduk ve bilgi sizintisini önlemek amacıyla onu bzersiz tanımlayıcılarla (grafik kimliği ve hikaye kimliği) göre eğitim ve test kümelerine böldük. DeepSeek-R1'i (DeepSeek-AI 2025) kullanarak (sıcaklık=0.6, maksimum jeton=8192, ayrintılar Ek A.3'te), CausalDR veri kümesi için yüksek kaliteli eğitim örnekleri oluşturmak ve doğrulamak amacıyla otomatik bir boru hattı (sözde kod Ek A.5'te verilmiştir) geliştirdik. Özellikle, DeepSeek-R1'e her CLADDER örneği için JSON formatında çıktılar üretmesini talimat veren, biçimlendirme matematiksel olarak doğru bir nedensel YADÇ içeren bir komut (ayrintılar Ek A.4'te) tasarladık. Her çıktı açıkça şunları içerir: (1) net semantik açıklamalara sahip nedensel düğümler ve gelen-giden ilişkileri belirten nedensel kenarlar, (2) oluşturulan nedensel YADÇ'ye açıkça atıfta bulunan adım adım bir akıl yürütme yolu ve (3) nihai çıkarılan cevabı. Kaliteyi garanti etmek amacıyla, model tarafından üretilen cevapları CLADDER tarafından sağlanan orijinal zemin gerceği cevaplarıyla karşılaştırın bir doğrulama mekanizması uyguladık. Birden fazla doğrulama denemesinin ardından üretilen bir cevap zemin gerceği ile eşleşmediğinde, örnek manuel olarak incelendi veya elendi.

Bu süreçle 6.357 doğrulanmış örnekten oluşan yüksek kaliteli bir veri kümesi elde ettik.

Veri kümesi çeşitliliğini ve genelleştirilebilirliğini daha da artırmak amacıyla bir Nedensel YADÇ Artırma teknigi kullandık. Özellikle, orijinal bir nedensel YADÇ $G = (V, E)$ verildiğinde, sırasıyla $\pi_V(\cdot)$ ve $\pi_E(\cdot)$ permütasyon fonksiyonlarını kullanarak nedensel düğümlerin ve kenarların sırasını rastgele değiştirdik ve böylece çeşitli artırılmış varyantlar oluşturduk: $V_{\text{aug}} = \pi_V(V)$, $E_{\text{aug}} = \pi_E(E)$, $G_{\text{aug}} = (V_{\text{aug}}, E_{\text{aug}})$. Bu permütasyon prosedürü orijinal her bir DAG G için dört kez uyguladık; her seferinde permüte edilmiş DAG G_{aug} 'i orijinal muhakeme yolu P ve yanıt y ile eşleştirildi. Bu, başlangıçtaki 6.357 örnekten oluşan veri kümescini genişletti.

örnek sayısını 25.368 artırılmış eğitim örnekine çıkardı. Şekil 4, CausalDR veri kümesindeki 12 kanonik nedensel YADÇ yapısının oransal dağılımını göstermektedir.

Bu yapılar, basit Zincirler (örn. Zincir, Çift Zincir), Karıştırıcı yapıları (örn. Üçgen, Çatal), Çarpıştırıcılar (örn. Çarpıştırıcı, Çoklu Çarpıştırıcı) ve daha karmaşık çok yolu etkileşimler (örn. Elmas, Genişletilmiş Papyon) dahil olmak üzere çeşitli nedensel konfigürasyonları kapsar. Bu temel nedensel mekanizmaların çeşitli temsili, büyük dil modellerinde nedensel akıl yürütme yeteneklerinin etkili bir şekilde genelleştirilmesini sağlar.

Yardımcı Talimatları Takip Eden Veriler Modelin nedensel görevde aşırı odaklanmasını ve dilsel genelleme yeteneğinin bozulmasını önlemek amacıyla, 10.000 Alpaca (Taori ve diğerleri 2023) örneğini rastgele seçip denetimli ince ayar sırasında CausalDR veri kümesi ile karıştırarak modelin genel dilsel yeteneğini ve genelleme performansını sağlamaktayız

3.3 Denetimli İnce Ayar Prosedürü

Denetimli ince ayar sırasında, BDM, yapılandırılmış nedensel Yönlendirilmiş Döngüsel Çizge çıkışım dizisini oluşturmayı öğrenir: $S_{\text{CDCR-SFT}} = (G, P, y)$. Optimizasyon hedefi, negatif log-olasılık kaybı olarak formüle edilmiştir: $\mathcal{L}_{\text{CDCR-SFT}} = -\sum_{t=1}^{|S|} \log P(s_t | s_{<t}, X)$; burada s_t , zemin gerçeki dizisindeki t -inci tokeni ve $s < t$ ise t konumundan önceki tüm tokenları temsil eder.

Kritik olarak, model tarafından oluşturulan diziler, zemin gerceği nedensel Yönlendirilmiş Döngüsel Olmayan Çizge yapısından saplığında (örneğin, ters nedensel kenarların tanıtıması, temel nedensel değişkenlerin atlanması veya fazla dan nedensel ilişkilerin eklenmesi gibi durumlarda), açık gradyan sinyalleri bu yanlışlıklarını aradık. Bu denetim, modelin doğru nedensel akıl yürütme için gerekli olan nedensel yönülüyü, koşullu bağımsızlık özelliklerini ve müda-hale semantığını doğru bir şekilde içselleştirmesini sağlar

Hesaplama verimliliği için, ince ayar sırasında Düşük Dereceli Adaptasyon (LoRA) (Hu ve diğerleri 2022) uyguladık. Bu süreçte, orijinal önceden eğitilmiş BDM parametrelerini dondururken, her katmana eklenen az sayıda düşük dereceli parametre güncellendik. Bu ince ayar prosedürü aracılığıyla, CDCR-SFT, modeli doğru bir nedensel Yönlendirilmiş Döngüsel Olmayan Çizge (YADÇ) oluşturması ve bu çizgelerde tanımlanan nedensel ilişkilerle açıkça kısıtlanmış yapılandırılmış akıl yürütme gerçekleştirmesi için eğitir. Böylece, BDM çıktılarının mantıksal tutarlılığı artırılır ve halüsinsiyonlar azaltılır

4 Deneyler

4.1 Deneysel Kurulum

Temel BDM'ler ve Akıl Yürütme Yöntemleri Değerlendirme için dört adet önceden eğitilmiş BDM seçti: (1) Llama-3.1-8B-Instruct (Grattafiori ve diğerleri 2024), (2) DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI 2025), (3) Baichuan2-7B-Chat (Baichuan 2023) ve (4) Mistral-7B-Instruct-v0.2 (Jiang ve diğerleri 2023). Ek olarak, CDCR-SFT yöntemimizi beş temel akıl yürütme yaklaşımıyla karşılaştırıyoruz: Sıfır Atılıcı Düşünce Zinciri (Düşünce Zinciri) (Kojima vd. 2023), Düşünce Zinciri Kendi Tutarlılığı (Düşünce Zinciri-SC) (Wang vd.

2023), Causal Chain-of-Thought (**CausalCoT**) (Jin et al. 2023), Tree-of-Thoughts (**ToT**) (Yao et al. 2023), and Graph-of-Thoughts (**GoT**) (Besta et al. 2024). Detailed descriptions of each baseline are provided in Appx. B.1.

Datasets We conduct experiments on 3 distinct datasets (Cladder (Jin et al. 2023), WIQA (Tandon et al. 2019), and HaluEval (Li et al. 2023)) to evaluate models’ causal reasoning and hallucinations performance.

Dataset	#	Subtasks	Evaluation Focus
CLADDER	1,906	Rung 1, Rung 2, Rung 3	Causal reasoning; Causal DAG quality
WIQA	212	INPARA, EXOGENOUS	Causal reasoning
HaluEval	1,500	Dialogue, QA, Summarization	Hallucination

Table 2: Summary of datasets used in experiments.

CLADDER (Jin et al. 2023): A benchmark dataset evaluating LLMs’ causal reasoning at three levels: Rung 1 (Association, observational correlations), Rung 2 (Intervention, active manipulation effects), and Rung 3 (Counterfactual, hypothetical “what-if” scenarios). Following preprocessing (see section 3.2), CLADDER is split into training and test sets by `graph_id` and `story_id` to avoid data leakage. To further ensure test data quality, we perform an additional validation step (details in Appx. B.2).

WIQA (Tandon et al. 2019): A challenging dataset for evaluating LLMs’ causal reasoning capabilities. We focus on two perturbation types: in-paragraph (INPARA)—changes within the text that test causal chain reconstruction, and out-of-paragraph (EXOGENOUS)—external changes assessing the model’s reasoning about external influences. We excluded irrelevant (no-effect) perturbations, as these modifications are unrelated to the original causal chain and do not effectively reflect a model’s true causal reasoning capability. For efficiency and representativeness, we sampled 106 questions per subtask (total 212; 95% confidence, $\pm 7\%$ margin). Questions were systematically reformulated to reduce ambiguity (see Appx. B.3).

HaluEval (Li et al. 2023): A benchmark for evaluating models’ hallucination across three NLP tasks: (1) Knowledge-grounded Dialogue (Dialogue), (2) Question Answering (QA), and (3) Text Summarization (Summarization). Each task includes paired examples, consisting of hallucinated samples (incorrect or unverifiable information) and corresponding factual samples. For our experiments, we randomly sample 500 pairs per task (total 1,500 pairs).

Evaluation Metrics We adopt two primary metrics to clearly evaluate the models’ causal reasoning and hallucination reduction: (1) **Accuracy**: measures correctness in causal reasoning (CLADDER, WIQA) and hallucination (HaluEval) tasks. (2) **Causal DAG Quality**: evaluates *Node Score* (correct causal nodes), *Edge Score* (correct causal edges), and *Structural Score* (overall graph correctness, including directionality and completeness). Causal DAG is scored using GPT-4o-mini (Hurst et al. 2024), with detailed scoring criteria and evaluation procedures provided in Appx. B.4.

Implementation Details We perform LoRA fine-tuning on A40x4 GPUs using the LLaMA-Factory library (Zheng et al. 2024) with default hyperparameters. Fine-tuned models use vLLM (Kwon et al. 2023) on the same GPUs for inference. Base model inference is conducted through external platforms: DeepInfra for Llama-3.1-8B and Mistral-7B, and Baidu-Qianfan/OpenRouter for Baichuan2-7B and DeepSeek-R1-Distill-Llama-8B, with 200 concurrent threads. The inference temperature is set to 0.0 except for DeepSeek (0.6, following (DeepSeek-AI 2025)), CoT-SC (0.7, following (Wang et al. 2023)), and GoT (1.0, following (Besta et al. 2024)). Our method and CoT-based approaches utilize a unified three-step instruction, while CausalCoT, ToT, and GoT follow their original structured prompting (Jin et al. 2023; Yao et al. 2023; Besta et al. 2024). All reported results are averaged over 3 experimental runs. Complete implementation details, prompts, and configurations are provided in our available code.

4.2 Main Results and Analysis

Causal Reasoning Performance Table 3 reports the causal reasoning performance of our proposed CDCR-SFT method compared with five baseline methods (CoT, CoT-SC, CausalCoT, ToT, and GoT) across four different LLMs on two representative causal reasoning benchmarks: CLADDER and WIQA.

On the CLADDER benchmark, our CDCR-SFT consistently achieves improvements across all three causal reasoning levels (Rung 1: Association, Rung 2: Intervention, and Rung 3: Counterfactual). Specifically, with the Llama-3.1-8B-Instruct model, our method reaches an overall accuracy of 95.33%, surpassing the strongest baseline (CoT-SC: 72.88%) by an absolute margin of 22.45 percentage points. Remarkably, at the most challenging Counterfactual reasoning level (Rung 3), CDCR-SFT achieves a particularly large improvement of 27.75 percentage points, improving accuracy from 65.31% (CoT-SC) to 93.06%. More importantly, our approach is the first to surpass the human-level benchmark performance (94.8%) (Yu et al. 2025) on CLADDER. Similar consistent performance gains are also observed for the DeepSeek-R1-Distill-Llama-8B (74.29% to 92.44%), Baichuan2-7B-Chat (52.26% to 72.51%), and Mistral-7B-Instruct-v0.2 (59.60% to 92.76%) models.

On the WIQA benchmark, CDCR-SFT again achieves consistent improvements over all baseline methods. Taking the Llama-3.1-8B-Instruct model as an example, the overall accuracy is improved from the best baseline (CoT-SC: 52.36%) to 55.66%. Similar improvements are consistently observed for the DeepSeek-R1-Distill-Llama-8B (52.83% to 55.66%), Baichuan2-7B-Chat (33.49% to 50.00%), and Mistral-7B-Instruct-v0.2 (41.51 to 44.81%) models.

These consistent gains across multiple causal reasoning tasks (CLADDER and WIQA) and diverse model architectures—from instruction-tuned models (Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.2) to distilled variants (DeepSeek-R1-Distill-Llama-8B) and smaller-scale models (Baichuan2-7B-Chat)—reflect that the benefits of CDCR-SFT originate primarily from its explicit modeling of causal structures reasoning. Unlike conventional meth-

2023), Nedensel Düşünce Zinciri (**CausalCoT**) (Jin vd. 2023), Düşünce Ağacı (**ToT**) (Yao vd. 2023) ve Düşünce Çizgesi (**GoT**) (Besta ve diğerleri. 2024). Her bir temel doğrulama yönteminin ayrıntılı açıklamaları Ek B.1’de sunulmuştur.

Veri Kümeleri Modellerin nedensel akıl yürütme ve halüsinasyon performansını değerlendirmek amacıyla 3 farklı veri kümesi (Cladder (Jin vd. 2023), WIQA (Tandon vd. 2019) ve HaluEval (Li vd. 2023)) üzerinde deneyler yapıyoruz.

Veri Kümesi	#	Alt Görevler	Değerlendirme Odağı
CLADDER	1,906	Basamak 1, Basamak 2, Basamak 3 Nedensel akıl yürütme ; Nedensel YADC kalitesi	
WIQA	212	INPARA, EKSOJEN	Nedensel akıl yürütme
HaluEval	1,500	Diyalog, KA, Özetteleme	Halüsinasyon

Tablo 2: Deneylerde kullanılan veri setlerinin özeti.

CLADDER (Jin vd. 2023): Büyük Dil Modellerinin nedensel akıl yürütmesini üç seviyede değerlendiren bir karşılaştırma veri seti: Basamak 1 (İlişkilendirme, gözlemsel korelasyonlar), Basamak 2 (Müdahale, aktif manipülasyon etkileri) ve Basamak 3 (Karşıolgusal, hipotetik “eğer olsaydı ne olurdu” senaryoları). Ön işlemeyi takiben (bkz. bölüm 3.2), veri sizintisini önlemek amacıyla CLADDER, grafik kimliği ve hikaye kimliği ile eğitim ve test setlerine ayrılır. Test verisi kalitesini daha da sağlamak için ek bir doğrulama adımı gerçekleştiriyoruz (ayrintılar Ek B.2’de).

WIQA (Tandon vd. 2019): Büyük Dil Modellerinin nedensel akıl yürütme yeteneklerini değerlendirmek için zorlayıcı bir veri seti. İki tür perturbasyona odaklıyoruz: paragraf içi (INPARA)—nedensel zincir yeniden yapılandırmasını test eden metin içindeki değişiklikler ve paragraf dışı (EKSOJEN)—modelin dış etkiler hakkındaki akıl yürütmesini değerlendiren harici değişiklikler. İlgisiz (etkisiz) perturbasyonları hariç tuttuk, çünkü bu modifikasyonlar orijinal nedensel zincirle alakasızdır ve bir modelin gerçek nedensel akıl yürütme yeteneğini etkili bir şekilde yansıtmadır. Verimlilik ve temsil yeteneği için, her alt görevden 106 soru örnekledik (toplam 212; %95 güven, $\pm 7\%$ hata payı). Belirsizliği azaltmak amacıyla sorular sistematik olarak yeniden formül edildi (Bkz. Ek B.3).

HaluEval (Li vd. 2023): Üç NLP görevinde modellerin halüsinasyonlarını değerlendirmek için bir kıyaslama: (1) Bilgiye Dayalı Diyalog (Diyalog), (2) Soru Cevaplama (KA) ve (3) Metin Özetteleme (Özetteleme). Her görev, halüsinasyonlu örnekler (yanlış veya doğrulanamayan bilgi) ve bunlara karşılık gelen gerçek örneklerden oluşan eşleştirilmiş örnekler içerir. Deneylerimiz için, görev başına rastgele 500 çift örnekledik (toplam 1.500 çift).

Değerlendirme Metrikleri Modellerin nedensel akıl yürütmesini ve halüsinasyon azaltımını açıkça değerlendirmek için iki temel metrik benimsiyor: (1) Doğruluk : nedensel akıl yürütme (CLADDER, WIQA) ve halüsinasyon (HaluEval) görevlerindeki doğruluğu ölçer. (2) Nedensel Yönlendirilmiş Döngüsel Olmayan Çizelge Kalitesi : Düğüm Skoru (doğru nedensel düğümler), Kenar Skoru (doğru nedensel kenarlar) ve Yapisal Puanı (yon ve eksiksizlik dahil olmak üzere genel çizelge doğruluğu) değerlendirir. Nedensel Yönlendirilmiş Döngüsel Olmayan Çizelge, GPT-4o-mini (Hurst vd. 2024) kullanılarak puanlanır; detaylı puanlama kriterleri ve değerlendirme prosedürleri Ek B.4’te sunulmuştur.

Uygulama Detayları LLaMA-Factory kütüphanesi (Zheng vd. 2024) ile varsayılan hiperparametreler kullanarak A40x4 GPU’larda LoRA ince ayarını gerçekleştiriyoruz. İnce ayarlı modeller , çıkarım için aynı GPU'larda vLLM (Kwon vd. 2023) kullanır. Temel model çıkarımı, Llama-3.1-8B ve Mistral-7B için DeepInfra , Baichuan2-7B ve DeepSeek-R1-Distill-Llama-8B için Baidu-Qianfan/OpenRouter gibi harici platformlar aracılığıyla 200 eşzamanlı iş parçacığı ile yürütülür. Çıkarım sıcaklığı, DeepSeek (0.6, (DeepSeek-AI 2025) takip ederek), Düşünce Zinciri-SC (0.7, (Wang vd. 2023) takip ederek) ve Oyun Teorisi (1.0, (Besta ve diğerleri. 2024) takip ederek) hariç 0.0 olarak ayarlanmıştır. Yönettimiz ve Düşünce Zinciri tabanlı yaklaşımlar birleşik üç adımlı bir talimat kullanırken, CausalCoT, ToT ve Oyun Teorisi kendi orijinal yapılandırılmış yönlendirmelerini takip etmektedir (Jin vd. 2023; Yao vd. 2023; Besta ve diğerleri. 2024). Bildirilen tüm sonuçlar, 3 deneyel çalışma üzerinden ortalanmıştır. Tam uygulama detayları, istemler ve yapılandırmalar mevcut kodumuzda sağlanmıştır.

4.2 Ana Sonuçlar ve Analiz

Nedensel Akıl Yürütme Performansı Tablo 3, önerilen CDCR-SFT yöntemimizin nedensel akıl yürütme performansını, beş temel yöntemle (Düşünce Zinciri, Düşünce Zinciri-SC, CausalCoT, ToT ve Oyun Teorisi) karşılaştırarak, iki temsili nedensel akıl yürütme kıyaslama ölçüyü olan CLADDER ve WIQA üzerinde dört farklı BDM’de rapor etmektedir.

CLADDER karşılaştırma testinde, CDCR-SFT’imiz her üç nedensel akıl yürütme seviyesinde (Basamak 1: İlişkilendirme , Basamak 2: Müdahale ve Basamak 3: Karşıolgusal) tutarlı iyileştirmeler kaydetmektedir. Özellikle, Llama-3.1-8B-Instruct modeliyle, yöntemimiz %95,33’lük genel bir doğruluk oranına ulaşarak en güçlü temel çizgisi (CoT-SC: %72,88) mutlak 22,45 yüzde puanı farkla geride bırakmaktadır. Dikkate değer bir şekilde, en zorlu Karşıolgusal akıl yürütme seviyesinde (Basamak 3), CDCR-SFT, %65,31’den (CoT-SC) %93,06’ya yükselsek %27,75’lik özellikle büyük bir iyileşme sağlamamaktadır. Daha da önemlisi, yaklaşıımız CLADDER üzerindeki insan seviyesi karşılaştırma performansını (%94,8) (Yu vd. 2025) aşan ilk yöntemdir. Benzer tutarlı performans artıları DeepSeek-R1-Distill-Llama-8B (%74,29’dan %92,44’e), Baichuan2-7B-Chat (%52,26’dan %72,51’e) ve Mistral-7B-Instruct-v0.2 (%59,60’tan %92,76’ya) modelleri için de gözlemlenmiştir.

WIQA kıyaslama testinde, CDCR-SFT yine tüm temel yöntemlere göre tutarlı iyileştirmeler sağlamaktadır. Llama-3.1-8B-Instruct modeli örnek alındığında, genel doğruluk en iyi temel seviyeden (Düşünce Zinciri-SC: %52,36) %55,66’ya yükselmiştir. Benzer iyileştirmeler DeepSeek-R1-Distill-Llama-8B (%52,83’ten %55,66’ya), Baichuan2-7B-Chat (%33,49’dan %50,00’e) ve Mistral-7B-Instruct-v0.2 (%41,51’den %44,81’e) modelleri için de tutarlı bir şekilde gözlemlenmiştir.

Birden fazla nedensel akıl yürütme görevi (CLADDER ve WIQA) ve farklı model mimarileri—komut ayarlı modellerden (Llama-3.1-8B-Instruct ve Mistral-7B-Instruct-v0.2) damıtılmış varyantlara (DeepSeek-R1-Distill-Llama-8B) ve daha küçük ölçekli modellere (Baichuan2-7B-Chat) kadar—bu tutarlı kazanımlar, CDCR-SFT’nin faydalarının öncelikle nedensel yapıların akıl yürütülmesinin açıkça modellenmesinden kaynaklandığını yansımaktadır. Geleneksel yöntemlerin aksine-

Method	Cladder (%)↑			WIQA (%)↑			HaluEval (%)↑				
	Rung1	Rung2	Rung3	overall.	INPARA	EXOGENOUS	overall.	Dialogue	QA	Summarization	overall.
Llama-3.1-8B											
CausalCoT	70.90	72.82	57.46	65.90	48.11	33.96	41.04	56.40	42.60	56.20	51.73
CoT	69.07	82.06	57.33	66.95	54.72	45.28	50.00	50.60	39.80	55.60	48.67
CoT-SC	72.87	88.13	65.31	72.88	60.38	44.34	52.36	43.60	34.00	52.60	43.40
ToT	71.17	79.16	64.79	70.20	56.60	45.28	50.94	52.80	42.00	58.00	50.93
GoT	61.21	76.78	58.90	63.38	55.66	47.17	51.42	50.20	43.40	50.20	47.93
CDCR-SFT (Ours)	98.30	93.93	93.06	95.33	64.20	47.20	55.66	60.80	44.80	59.20	54.93
DeepSeek-R1-Distill-Llama-8B											
CausalCoT	74.97	68.87	59.03	67.37	52.83	50.94	51.89	47.60	40.00	51.80	46.47
CoT	73.92	76.78	53.27	66.21	55.66	47.17	51.42	42.00	40.80	48.00	43.60
CoT-SC	77.98	88.13	63.74	74.29	51.89	43.40	47.64	33.60	41.40	32.20	35.73
ToT	70.34	80.62	66.18	71.14	56.60	44.34	50.47	39.40	43.80	40.20	41.13
GoT	75.23	80.97	57.63	69.43	55.66	50.00	52.83	53.40	41.00	50.60	48.33
CDCR-SFT (Ours)	94.89	90.50	90.97	92.44	56.60	54.72	55.66	48.60	44.40	52.60	48.53
Baichuan2-7B											
CausalCoT	50.46	46.44	51.70	50.16	22.64	27.36	25.00	45.80	48.40	43.20	45.80
CoT	49.67	62.01	48.56	51.68	34.91	27.36	31.13	44.20	46.60	45.80	45.53
CoT-SC	51.38	61.21	48.69	52.26	36.79	30.19	33.49	47.80	45.80	47.40	47.00
ToT	49.67	58.05	50.65	51.73	34.91	20.75	27.83	44.80	45.80	48.01	46.20
GoT	51.11	58.84	49.61	52.05	31.13	30.19	30.66	41.80	43.80	40.80	42.13
CDCR-SFT (Ours)	71.04	75.20	72.64	72.51	50.00	50.00	50.60	49.60	51.00	50.40	
Mistral-7B											
CausalCoT	51.11	63.06	45.16	51.10	38.68	27.36	33.02	45.20	47.80	41.60	44.87
CoT	52.29	59.63	53.53	54.25	40.57	34.91	37.74	43.60	44.20	43.80	43.87
CoT-SC	56.75	66.75	58.90	59.60	42.45	38.68	40.57	44.40	45.20	44.00	44.53
ToT	50.46	56.20	50.39	51.57	42.45	32.08	37.26	47.00	42.80	46.60	45.47
GoT	50.85	63.85	56.15	55.56	42.45	40.57	41.51	47.60	46.20	46.80	46.87
CDCR-SFT (Ours)	94.23	94.46	90.45	92.76	43.40	46.23	44.81	53.40	48.20	53.60	51.73

Table 3: Performance comparison between our proposed CDCR-SFT and baseline reasoning methods on causal reasoning benchmarks (CLADDER and WIQA) and hallucination benchmark (HaluEval) across four different LLMs. Accuracy (%) is reported for overall benchmarks and subtasks; best results per model and task highlighted in bold.

ods that perform token-level or semantic-level reasoning, our approach trains LLMs to explicitly construct and reason over causal DAG, thus embedding a stronger inductive bias aligned with causal inference principles. Consequently, the models internalize improved representations of conditional independencies, intervention semantics, and causal directionality, facilitating more robust generalization across causal reasoning scenarios and tasks of varying complexity.

Hallucination Reduction Table 3 further reports the hallucination reduction performance of our proposed CDCR-SFT method across four different LLMs, evaluated on the HaluEval benchmark comprising three typical tasks: Dialogue, QA, and Summarization.

Our CDCR-SFT method consistently outperforms baseline reasoning methods in terms of overall accuracy on the HaluEval benchmark, demonstrating clear reductions in logical inconsistencies and hallucinations. Specifically, using the Llama-3.1-8B model, CDCR-SFT achieves an overall accuracy of 54.93%, significantly higher than the strongest baseline (CausalCoT: 51.73%) and substantially surpassing CoT-SC (43.40%) by over 11 percentage points. Particularly noteworthy is the Dialogue subtask, where accuracy improves from 43.60% (CoT-SC) to 60.80%, highlighting the effectiveness of our approach in mitigating hallucinations in complex interactive reasoning tasks.

Similar trends are evident for other evaluated LLMs. For

instance, the DeepSeek improves from the strongest baseline (CausalCoT: 46.47%) to 48.40%, Baichuan improves from 47.00% (CoT-SC) to 50.40%, and Mistral shows accuracy improvement from the best baseline (GoT: 46.87%) to 51.73%. Importantly, these significant hallucination reductions are achieved without hallucination-focused supervision, indicating that the reduction naturally arises from enhanced causal reasoning capabilities learned by the model.

These empirical findings directly support our core hypothesis: explicitly improving the causal reasoning capabilities of LLMs inherently mitigates logically inconsistent hallucinations. The substantial and consistent hallucination reductions observed across diverse tasks and model architectures demonstrate that our CDCR-SFT method provides an effective and generalizable solution for enhancing the reliability and consistency of LLMs.

4.3 Causal DAG Construction Quality

CDCR-SFT is to enable LLMs to reason accurately based on a variable-level causal DAG. The quality of the generated DAG thus directly reflects the extent to which the model has internalized correct causal relationships and structured causal reasoning capabilities, including accurately capturing causal directionality, conditional independencies, and satisfying causal identification assumptions. We compare the Causal DAG generated using pre-trained LLMs versus the

Yöntem	Cladder (%)↑			WIQA (%)↑			HaluEval (%)↑				
	Kademe1	Kademe2	Kademe3	genel.	INPARA	EKSÖJEN	genel .	Diyalog KA	Oztleme genel .		
Llama-3.1-8B											
CausalCoT	70.90	72.82	57.46	65.90	48.11	33.96	41.04	56.40	42.60	56.20	51.73
Düşünce Zinciri	69.07	82.06	57.33	66.95	54.72	45.28	50.00	50.60	39.80	55.60	48.67
Düşünce Zinciri-SC	72.87	88.13	65.31	72.88	60.38	44.34	52.36	43.60	44.34	52.60	43.40
ToT	71.17	79.16	64.79	70.20	56.60	45.28	50.94	52.80	42.00	58.00	50.93
Oyun Teorisi	61.21	76.78	58.90	63.38	55.66	47.17	51.42	50.20	43.40	50.20	47.93
CDCR-SFT (Bizimki)	98.30	93.93	93.06	95.33	64.20	47.20	55.66	60.80	44.80	59.20	54.93
DeepSeek-R1-Distill-Llama-8B											
CausalCoT	74.97	68.87	59.03	67.37	52.83	50.94	51.89	47.60	40.00	51.80	46.47
Düşünce Zinciri	73.92	76.78	53.27	66.21	55.66	47.17	51.42	42.00	40.80	48.00	43.60
Düşünce Zinciri-SC	77.98	88.13	63.74	74.29	51.89	43.40	47.64	33.60	41.40	32.20	35.73
ToT	70.34	80.62	66.18	71.14	56.60	44.34	50.47	39.40	43.80	40.20	41.13
Oyun Teorisi	75.23	80.97	57.63	69.43	55.66	50.00	52.83	55.66	41.00	50.60	48.33
CDCR-SFT (Bizimki)	94.89	90.50	90.97	92.44	56.60	54.72	55.66	56.60	44.40	52.60	48.53
Baichuan2-7B											
CausalCoT	50.46	46.44	51.70	50.16	22.64	27.36	25.00	45.80	48.40	43.20	45.80

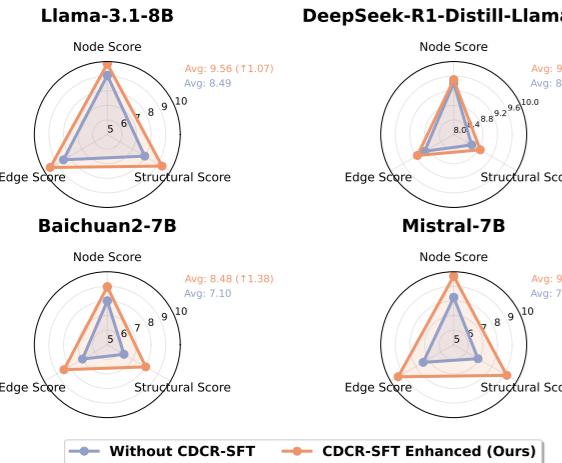


Figure 5: Comparison of causal DAG quality scores (Node, Edge, and Structural Scores) generated by pretrained LLMs versus those enhanced with CDCR-SFT, evaluated on the CLADDER dataset.

DAG produced by LLMs enhanced with our CDCR-SFT approach. Both employ the same prompt instructing models to generate Causal DAG. Fig. 5 indicates that CDCR-SFT raises scores in each dimension for all models. For Llama-3.1-8B, the overall average increases from 8.49 to 9.56, with the largest rise in Structural Score (7.96 to 9.33). DeepSeek-R1-Distill-Llama-8B shows a small increase from 8.97 to 9.17, chiefly in Edge Score (8.92 to 9.15). Baichuan2-7B advances from 7.10 to 8.48, with a 1.72-point gain in Structural Score. Mistral-7B displays the greatest progress, from 7.53 to 9.43, with gains over 2 points in both Edge Score and Structural Score. The significantly higher DAG quality achieved by CDCR-SFT over baseline methods validates that explicit DAG-based reasoning supervision enhances LLMs’ capability to correctly represent and reason with causal structures, directly supporting improvements observed in causal reasoning tasks and hallucination reduction.

4.4 Ablation Study

We conduct an ablation study to assess whether the observed performance improvements originate specifically from our causal DAG construction and causal DAG-based reasoning strategy, or merely from the additional exposure to causal knowledge and examples provided during fine-tuning. Specifically, we compare three experimental conditions across all three benchmarks, reporting overall accuracy for CLADDER, WIQA, and HaluEval: (i) *Baseline*: the best-performing existing reasoning method per benchmark (selected from CoT, CoT-SC, ToT, GoT, and CausalCoT in Table 3); (ii) *CDCR-SFT-Ablated*: fine-tunes LLMs using only question-answer pairs from the CausalDR dataset, omitting causal DAG G construction and reasoning paths P , but retaining identical auxiliary instruction-following data; (iii) *CDCR-SFT*: our full proposed method, explicitly trained on causal DAG construction and DAG-based reasoning. All conditions maintain identical training configu-

rations, including model architectures, hyperparameters, and data volumes, ensuring a fair comparison. Table 4 shows that

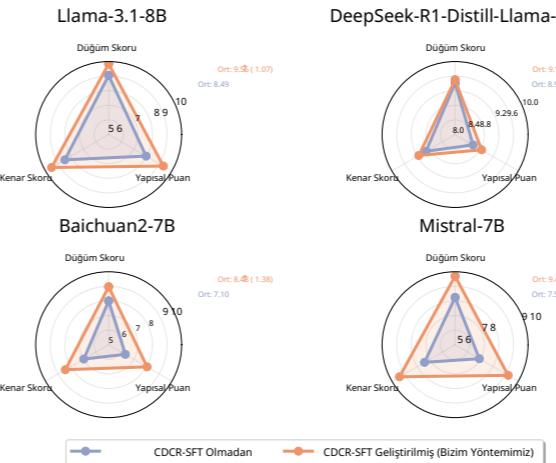
Method	Cladder (%)↑	WIQA (%)↑	HaluEval (%)↑
Llama-3.1-8B			
Baseline	72.88	52.36	51.73
CDCR-SFT-Ablated	87.25	49.06	44.97
CDCR-SFT (Ours)	95.33	55.66	54.93
DeepSeek-R1-Distill-Llama-8B			
Baseline	74.29	52.83	48.33
CDCR-SFT-Ablated	74.87	51.89	43.67
CDCR-SFT (Ours)	92.44	55.66	48.53
Baichuan2-7B			
Baseline	52.26	33.49	47.00
CDCR-SFT-Ablated	69.57	42.92	42.10
CDCR-SFT (Ours)	72.51	50.00	50.40
Mistral-7B			
Baseline	59.60	41.51	46.87
CDCR-SFT-Ablated	67.58	38.68	49.10
CDCR-SFT (Ours)	92.76	44.81	51.73

Table 4: Ablation study verifying the impact of explicit causal DAG-based reasoning, comparing baseline (best existing method), CDCR-SFT-Ablated (fine-tuned without causal DAG construction and reasoning), and our CDCR-SFT across three benchmarks on four LLMs.

fine-tuning models solely with causal question-answer pairs (CDCR-SFT-Ablated), without explicit causal DAG-based reasoning, consistently improves accuracy on the CLADDER benchmark (e.g., +14.4% on Llama-3.1-8B, +17.3% on Baichuan2-7B) but leads to performance degradation on the WIQA and HaluEval benchmarks compared to the Baseline. In contrast, our full method (CDCR-SFT), which learned causal DAG construction and causal DAG-based reasoning, consistently outperforms both the Baseline and CDCR-SFT-Ablated methods across all benchmarks and model architectures. These results confirm that the observed performance gains are attributable to structured causal reasoning rather than simply additional causal data exposure.

5 Conclusion

We propose CDCR-SFT, to shift how LLMs approach causal reasoning by moving from sequential CoT or graph variant to causal DAG-based reasoning. It trains models to construct a causal DAG that properly encodes both causal directionality and conditional independence relationships, enabling them to perform structured reasoning over the graph rather than being constrained by linear reasoning paths or causal-irrelevant graph reasoning. And we create the CausalDR dataset, containing 25,368 validated samples, provides high-quality supervision for LLMs to learn explicit causal DAG construction and graph-based reasoning. Our experiments across four LLMs on the CLADDER, WIQA, and HaluEval benchmarks demonstrate that CDCR-SFT significantly improves causal reasoning, achieving a state-of-the-art accuracy of 95.33% on CLADDER, surpassing human performance (94.8%) for the first time. Moreover, CDCR-SFT reduces hallucination in HaluEval by up to 11%, confirming that enhanced causal reasoning directly mitigates hallucinations. These results affirmatively



Şekil 5: Önceden eğitilmiş BDM'ler tarafından üretilen nedensel YADÇ kalite skorlarının (Dügüm, Kenar ve Yapısal Skorlar) CDCR-SFT ile geliştirilenlerle karşılaştırılması, CLADDER veri kümesi üzerinde de-gerlendirilmiştir.

adil bir karşılaştırma sağlar. Tablo 4, şunları göstermektedir:

Yöntem	Cladder (%)↑	WIQA (%)↑	HaluEval (%)↑
Llama-3.1-8B			
Temel Referans	72.88	52.36	51.73
CDCR-SFT-Ablated	87.25	49.06	44.97
CDCR-SFT (Bizim Yöntemimiz)	95.33	55.66	54.93
DeepSeek-R1-Distill-Llama-8B			
Temel Referans	74.29	52.83	48.33
CDCR-SFT-Ablated	74.87	51.89	43.67
CDCR-SFT (Bizim Yöntemimiz)	92.44	55.66	48.53
Baichuan2-7B			
Temel Referans	52.26	33.49	47.00
CDCR-SFT-Ablated	69.57	42.92	42.10
CDCR-SFT (Bizim Yöntemimiz)	72.51	50.00	50.40
Mistral-7B			
Temel Referans	59.60	41.51	46.87
CDCR-SFT-Ablated	67.58	38.68	49.10
CDCR-SFT (Bizim Yöntemimiz)	92.76	44.81	51.73

Tablo 4: Açık nedensel Yönlendirilmiş Dönüşel Olmayan Çizelge (YADÇ) tabanlı akıl yürütmenin etkisini doğrulan ablasyon çalışması; temel referans (mevcut en iyi yöntem), CDCR-SFT-Ablated (nedensel YADÇ inşası ve akıl yürütmesi olmadan ince ayar yapılmış) ve bizim CDCR-SFT yöntemimizi dört Büyük Dil Modeli (BDM) üzerinde üç kıyaslama ölçütünde karşılaştırmaktadır.

Modelleri yalnızca nedensel soru-cevap çiftleriyle (CDCR-SFT-Ablated) ve açık nedensel YADÇ tabanlı akıl yürütme olmaksızın ince ayar yapmak, CLADDER kıyaslama ölçütünde doğruluğu sürekli olarak artırırken (örn. Llama-3.1-8B'de +%14.4, Baichuan2-7B'de +%17.3), WIQA ve HaluEval kıyaslama ölçütlerinde Temel Referans'a kıyasla performans düşüşüne yol açmaktadır. Buna karşılık, nedensel YADÇ inşasını ve nedensel YADÇ tabanlı akıl yürütme öğrenen tam yöntemimiz (CDCR-SFT), tüm kıyaslama ölçütlerinde ve model mimarilerinde hem Temel Referans hem de CDCR-SFT-Ablated yöntemlerinden sürekli olarak daha iyi performans sergilemektedir. Bu sonuçlar, gözlemlenen performans artışılarının sadece ek nedensel veri maruziyetinden ziyade yapılandırılmış nedensel akıl yürütmeye atfedilebilir olduğunu doğrulamaktadır.

5 Sonuç

Büyük Dil Modellerinin (BDM) nedensel akıl yürütmeye yaklaşımını sıralı Düşünce Zinciri (CoT) veya grafik varyantından nedensel Yönlendirilmiş Dönüşel Çizge (YADÇ) tabanlı akıl yürütmeye kaydırma amacıyla CDCR-SFT'yi öneriyoruz. Bu çerçeve, modelleri hem nedensel yönlülüğü hem de koşullu bağımsızlık ilişkilerini doğru bir şekilde kodlayan nedensel YADÇ oluşturmalrı için eğitir; bu sayede modellerin, doğrusal akıl yürütme yolları veya nedensel olarak ilgiz grafik akıl yürütmeyle kısıtlanmak yerine, çizge üzerinde yapılandırılmış akıl yürütme gerçekleştirmelerini sağlar. Ayrıca, 25.368 doğrulanmış örnek içeren CausalDR veri kümescini oluşturdu; bu veri kümesci, BDM'lerin açık nedensel YADÇ inşası ve grafik tabanlı akıl yürütme öğrenmeleri için yüksek kaliteli denetim sağlar. Dört BDM üzerinde CLADDER, WIQA ve HaluEval kıyaslama testlerinde yaptığımız deneyler, CDCR-SFT'nin nedensel akıl yürütme önemde iyileştirdiğini, CLADDER üzerinde %95,33'lük son teknoloji doğruluğu ulaşarak insan performansını (%94,8) ilk kez geride bıraktığını göstermek. Dahası, CDCR-SFT, HaluEval'deki halüsinasyonları %11'e kadar azaltarak, gelişmiş nedensel akıl yürütmenin halüsinasyonları doğrudan hafiflettiğini doğrulamaktadır. Bu sonuçlar olumlu bir şekilde

: BDM'ler yalnızca CausalDR veri kümescinden gelen soru-cevap çiftlerini kullanarak ince ayarlar,

answer our research question: **improving the causal reasoning capabilities of LLMs can mitigate hallucinations**. In the future, rather than solely pursuing larger model sizes or more training data or longer cot, we can achieve more trustworthy LLMs by equipping them with structured reasoning capabilities that align with the underlying causal nature of real-world problems.

References

- Bagheri, A.; Alinejad, M.; Bello, K.; and Akhondi-Asl, A. 2024. C²P: Featuring Large Language Models with Causal Reasoning. *arXiv*:2407.18069.
- Baichuan. 2023. Baichuan 2: Open Large-scale Language Models. *arXiv preprint arXiv*:2309.10305.
- Banerjee, S.; Agarwal, A.; and Singla, S. 2024. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv*:2409.05746.
- Bao, G.; Zhang, H.; Wang, C.; Yang, L.; and Zhang, Y. 2024. How Likely Do LLMs with Cot Mimic Human Reasoning? *arXiv preprint arXiv*:2402.16048.
- Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawska, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the CDCR conference on artificial intelligence*, volume 38, 17682–17690.
- Cheng, F.; Li, H.; Liu, F.; van Rooij, R.; Zhang, K.; and Lin, Z. 2025. Empowering llms with logical reasoning: A comprehensive survey. *arXiv preprint arXiv*:2502.15652.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv*:2501.12948.
- Du, L.; Ding, X.; Xiong, K.; Liu, T.; and Qin, B. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv*:2205.05849.
- Fu, J.; Ding, L.; Li, H.; Li, P.; Wei, Q.; and Chen, X. 2025. Unveiling and causalizing cot: A causal pespective. *arXiv preprint arXiv*:2502.18239.
- Gordon, A.; Kozareva, Z.; and Roemmele, M. 2012. SemEval-2012 Task 7: Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. 394–398. Montréal, Canada: Association for Computational Linguistics.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv*:2407.21783.
- Han, Z.; Gao, C.; Liu, J.; Zhang, J.; and Zhang, S. Q. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv*:2403.14608.
- Hernan, M.; and Robins, J. 2020. Causal inference: What if Chapman Hall/CRC, Boca Raton.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv*:2410.21276.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *arXiv*:2310.06825.
- Jin, Z.; Chen, Y.; Leeb, F.; Greselle, L.; Kamal, O.; Lyu, Z.; Blin, K.; Gonzalez, F.; Kleiman-Weiner, M.; Sachan, M.; and Schölkopf, B. 2023. CLadder: Assessing Causal Reasoning in Language Models. In *NeurIPS*.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. *arXiv*:2205.11916.
- Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Li, J.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models.
- Liu, X.; Xu, P.; Wu, J.; Yuan, J.; Yang, Y.; Zhou, Y.; Liu, F.; Guan, T.; Wang, H.; Yu, T.; et al. 2025. Large language models and causal inference in collaboration: A comprehensive survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, 7668–7684.
- Luo, H.; Zhang, J.; and Li, C. 2025. Causal graphs meet thoughts: Enhancing complex reasoning in graph-augmented llms. *arXiv preprint arXiv*:2501.14892.
- Ma, J. 2024. Causal inference with large language model: A survey. *arXiv preprint arXiv*:2409.09822.
- Tandon, N.; Mishra, B. D.; Sakaguchi, K.; Bosselut, A.; and Clark, P. 2019. Wiqa: A dataset for “what if...” reasoning over procedural text. *arXiv preprint arXiv*:1909.04739.
- Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv*:2203.11171.
- Wang, Z. 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)*, 143–151. Bangkok, Thailand: Association for Computational Linguistics.
- araştırma sorumuzu yanıtlamaktadır: BDM'lerin nedensel akıl yürütme yeteneklerini geliştirmek halüsinasyonları hafifletebilir. Gelecekte, yalnızca daha büyük model boyutları veya daha fazla eğitim verisi ya da daha uzun düşünce zinciri peşinde koşmak yerine, BDM'leri gerçek dünya problemlerinin altında yatan nedensel doğasıyla uyumlu yapılandırılmış akıl yürütme yetenekleriyle donatarak daha güvenilir BDM'ler elde edilebilir.
- Referanslar**
- Bagheri, A.; Alinejad, M.; Bello, K.; ve Akhondi-Asl, A. 2024. C²P: Büyük Dil Modellerini Nedensel Akıl Yürütme ile Sunmak. *arXiv*:2407.18069.
- Baichuan. 2023. Baichuan 2: Açık Büyük Ölçekli Dil Modelleri. *arXiv ön baskısı arXiv*:2309.10305.
- Banerjee, S.; Agarwal, A.; ve Singla, S. 2024. BDM'ler her zaman halüsinasyon üretmeyecek ve bununla yaşamamız gerekecek. *arXiv ön baskısı arXiv*:2409.05746.
- Bao, G.; Zhang, H.; Wang, C.; Yang, L.; ve Zhang, Y. 2024. Düşünce Zinciri ile BDM'ler İnsan Akıl Yürütmesini Ne Kadar Taklit Eder? *arXiv ön baskısı arXiv*:2402.16048.
- Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawska, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; vd. 2024. Düşünce grafiği: Büyük dil modelleriyle karmaşık sorunları çözmeye. Yapay zeka üzerine CDCR konferansı bildirilerinde, cilt 38, 17682–17690.
- Cheng, F.; Li, H.; Liu, F.; van Rooij, R.; Zhang, K.; ve Lin, Z. 2025. BDM'leri mantıksal akıl yürütme ile güçlendirmek: Kapsamlı bir araştırma. *arXiv ön baskısı arXiv*:2502.15652.
- DeepSeek-AI. 2025. DeepSeek-R1: Takviyeli Öğrenme Yoluyla BDM'lerde Akıl Yürütme Yeteneğini Teşvik Etmek. *arXiv*:2501.12948.
- Du, L.; Ding, X.; Xiong, K.; Liu, T.; ve Qin, B. 2022. e-CARE: açıklanabilir nedensel akıl yürütme keşfetmek için yeni bir veri kümlesi. *arXiv ön baskısı arXiv*:2205.05849.
- Fu, J.; Ding, L.; Li, H.; Li, P.; Wei, Q.; ve Chen, X. 2025. Düşünce Zinciri'nin ortaya çıkarılması ve nedenselleştirilmesi: Nedensel bir bakış açısı. *arXiv ön baskısı arXiv*:2502.18239.
- Gordon, A.; Kozareva, Z.; ve Roemmele, M. 2012. SemEval-2012 Görev 7: Makul Alternatiflerin Seçimi: Sağıduyu Nedensel Akıl Yürütmenin Değerlendirilmesi. 394–398. Montréal, Kanada: Hesaplama Dil Bilimleri Derneği Buluşları: NAACL 2025 , 7668–7684.
- Luo, H.; Zhang, J.; ve Li, C. 2025. Nedensel çizgeler düşüncelerle buluşuyor: Çizge destekli BDM'lerde karmaşık akıl yürütme geliştirme. *arXiv ön baskısı arXiv*:2501.14892.
- Ma, J. 2024. Büyük dil modeliyle nedensel çıkarım: Bir anket. *arXiv ön baskısı arXiv*:2409.09822.
- Tandon, N.; Mishra, B. D.; Sakaguchi, K.; Bosselut, A.; ve Clark, P. 2019. Wiqa: Prosedürel metin üzerinde “ne olurdu...” akıl yürütmesi için bir veri kümlesi. *arXiv ön baskısı arXiv*:1909.04739.
- Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; ve Hashimoto, T. B. 2023. Stanford Alpaca: Talimatları takip eden bir LLaMA modeli. https://github.com/tatsu-lab/stanford_alpaca.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; ve Zhou, D. 2023. Dil Modellerinde Düşünce Zinciri Akıl Yürütmesini Geliştiren Öz-Tutarlılık. *arXiv*:2203.11171.
- Wang, Z. 2024. CausalBench: Büyük Dil Modellerinin Nedensel Akıl Yürütme Yeteneklerini Değerlendirmek İçin Kapsamlı Bir Kiyaslama. İçinde 10. SIGHAN Çince Dil İşleme Çalıştayı Bildirileri (SIGHAN-10) , 143–151. Bangkok, Tayland: Hesaplama Dil Bilimleri Derneği.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; vd. 2025. Büyük Dil Modellerinde halüsinasyon üzerine bir inceleme: İlkeler, taksonomi, zorluklar ve açık sorular. *ACM Bilgi Sistemleri Üzerine İşlemler* , 43(2): 1–55.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; ve diğerleri. 2024. Gpt-4o sistem kartı. *arXiv ön baskısı arXiv*:2410.21276 .
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; ve Sayed, W. E. 2023. Mistral 7B. *arXiv*:2310.06825.
- Jin, Z.; Chen, Y.; Leeb, F.; Greselle, L.; Kamal, O.; Lyu, Z.; Blin, K.; Gonzalez, F.; Kleiman-Weiner, M.; Sachan, M.; ve Schölkopf, B. 2023. CLadder: Dil Modellerinde Nedensel Akıl Yürütme Değerlendirme. In *NeurIPS* . Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; ve Iwasawa , Y. 2023. Büyük Dil Modelleri Sıfır Atılı Akıl Yürütürülerdir. *arXiv*:2205.11916.
- Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; ve Stoica, I. 2023. Pa gedAttention ile Büyük Dil Modeli Sunumu İçin Verimli Bellek Yönetimi. In *ACM SIGOPS 29. İşletim Sistemleri İlkeleri Sempozyumu Bildirileri* .
- Li, J.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; ve Wen, J.-R. 2025. HaluEval: Büyük Dil Modelleri İçin Büyük Ölçekli Bir Halüsinasyon Değerlendirme Kiyaslaması.
- Liu, X.; Xu, P.; Wu, J.; Yuan, J.; Yang, Y.; Zhou, Y.; Liu, F.; Guan, T.; Wang, H.; Yu, T.; vd. 2025. Büyük dil modelleri ve nedensel çıkarım işbirliği içinde: Kapsamlı bir araştırma. *Hesaplama Dil Bilimleri Derneği Buluşları: NAACL 2025* , 7668–7684.
- Luo, H.; Zhang, J.; ve Li, C. 2025. Nedensel çizgeler düşüncelerle buluşuyor: Çizge destekli BDM'lerde karmaşık akıl yürütme geliştirme. *arXiv ön baskısı arXiv*:2501.14892 .
- Ma, J. 2024. Büyük dil modeliyle nedensel çıkarım: Bir anket. *arXiv ön baskısı arXiv*:2409.09822.
- Tandon, N.; Mishra, B. D.; Sakaguchi, K.; Bosselut, A.; ve Clark, P. 2019. Wiqa: Prosedürel metin üzerinde “ne olurdu...” akıl yürütmesi için bir veri kümlesi. *arXiv ön baskısı arXiv*:1909.04739.
- Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; ve Hashimoto, T. B. 2023. Stanford Alpaca: Talimatları takip eden bir LLaMA modeli. https://github.com/tatsu-lab/stanford_alpaca.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; ve Zhou, D. 2023. Dil Modellerinde Düşünce Zinciri Akıl Yürütmesini Geliştiren Öz-Tutarlılık. *arXiv*:2203.11171.
- Wang, Z. 2024. CausalBench: Büyük Dil Modellerinin Nedensel Akıl Yürütme Yeteneklerini Değerlendirmek İçin Kapsamlı Bir Kiyaslama. İçinde 10. SIGHAN Çince Dil İşleme Çalıştayı Bildirileri (SIGHAN-10) , 143–151. Bangkok, Tayland: Hesaplama Dil Bilimleri Derneği.

- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.
- Yu, L.; Chen, D.; Xiong, S.; Wu, Q.; Liu, Q.; Li, D.; Chen, Z.; Liu, X.; and Pan, L. 2025. CausalEval: Towards Better Causal Reasoning in Language Models. arXiv:2410.16676.
- Zhang, Y.; Yuan, Y.; and Yao, A. C.-C. 2024. On the diagram of thought. arXiv preprint arXiv:2409.10038.
- Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Bangkok, Thailand: Association for Computational Linguistics.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; vd. 2022. Düşünce zinciri yönlenmesi büyük dil modellerinde akıl yürütme tetikler. *Sınırsız bilgi işleme sistemlerindeki gelişmeler*, 35: 24824–24837.
- Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; ve Narasimhan, K. 2023. Düşünce Ağacı: Büyük Dil Modelleriyle Özenli Problem Çözme. *Sınırsız Bilgi İşleme Sistemlerindeki Gelişmeler*, 36: 11809–11822.
- Yu, L.; Chen, D.; Xiong, S.; Wu, Q.; Liu, Q.; Li, D.; Chen, Z.; Liu, X.; ve Pan, L. 2025. CausalEval: Dil Modellerinde Daha İyi Nedenlesel Akıl Yürütmeye Doğru. arXiv:2410.16676.
- Zhang, Y.; Yuan, Y.; ve Yao, A. C.-C. 2024. Düşünce Diagramı Üzerine. arXiv ön baskısı arXiv:2409.10038 .
- Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; ve Ma, Y. 2024. LlamaFactory: 100'den Fazla Dil Modelinin Birleşik Verimli İnce Ayarı. İçinde *Hesaplamalı Dilbilim Derneği'nin 62. Yıllık Toplantısı Bildirileri (Cilt 3: Sistem Gösterimleri)*. Bangkok, Tayland: Hesaplamalı Dilbilim Derneği.

Supplementary Material

A Additional Methodological Details

A.1 Example of the Cladde Dataset

Our CausalDR dataset for supervised fine-tuning is based on the publicly available dataset, Cladde. The Cladde dataset was initially proposed to evaluate causal reasoning capabilities of large language models (LLMs). The dataset provides scenarios, questions, formal symbolic reasoning steps, and answers. Here, we showcase only the attributes directly relevant to our data construction process.

Below is a representative example from the Cladde dataset:

Original Cladde Dataset Sample

Scenario & Question: Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships: Husband has a direct effect on wife and alarm clock. Wife has a direct effect on alarm clock. For husbands that don't set the alarm and wives that don't set the alarm, the probability of ringing alarm is 8%. For husbands that don't set the alarm and wives that set the alarm, the probability of ringing alarm is 54%. For husbands that set the alarm and wives that don't set the alarm, the probability of ringing alarm is 41%. For husbands that set the alarm and wives that set the alarm, the probability of ringing alarm is 86%. For husbands that don't set the alarm, the probability of alarm set by wife is 74%. For husbands that set the alarm, the probability of alarm set by wife is 24%. If we disregard the mediation effect through wife, would husband positively affect alarm clock?

Reasoning:

Let X = husband; V2 = wife; Y = alarm clock.

$X \rightarrow V2, X \rightarrow Y, V2 \rightarrow Y$

$$\begin{aligned} E[Y_{X=1, V2=0} - Y_{X=0, V2=0}] \\ \sum_{V2=v} P(V2 = v | X = 0) * [P(Y = 1 | X = 1, V2 = v) - P(Y = 1 | X = 0, V2 = v)] \\ P(Y = 1 | X = 0, V2 = 0) = 0.08 \\ P(Y = 1 | X = 0, V2 = 1) = 0.54 \\ P(Y = 1 | X = 1, V2 = 0) = 0.41 \\ P(Y = 1 | X = 1, V2 = 1) = 0.86 \\ P(V2 = 1 | X = 0) = 0.74 \\ P(V2 = 1 | X = 1) = 0.24 \\ 0.74 * (0.86 - 0.41) + 0.24 * (0.54 - 0.08) = 0.32 \\ 0.32 > 0 \end{aligned}$$

Answer: Yes

In this example, the scenario describes causal interactions among variables, the question focuses explicitly on causal effects, and the reasoning provides symbolic and probabilistic calculations. These symbolic representations serve as important references when identifying causal nodes and relationships for our structured data generation.

A.2 Example of the CausalDR Dataset(Ours)

CausalDR dataset used for supervised fine-tuning within the CDCR-SFT framework is presented in a clear, structured JSON format. Each sample contains three essential components: (1) a specific **instruction** explicitly guiding the model on what kind of task it should perform, (2) an **input** that provides the detailed scenario description and the specific causal reasoning question, and (3) an **output** demonstrating a clear, structured, and comprehensive reasoning process explicitly based on the provided causal graph.

Below, we provide a detailed example from CausalDR dataset to illustrate precisely how we formulate these training samples. Each sample not only explicitly represents the causal graph, inference path, and answer but also cohesively integrates these components into a natural and intuitive explanatory narrative. This structured presentation is specifically designed to encourage LLMs to internalize a systematic approach to causal reasoning, thereby significantly enhancing inference consistency and reducing hallucinations.

Fine-tuning Dataset Example (JSON)

```
{  
  "instruction": "Given the question below, please construct a causal graph to analyze  
  the scenario. Then, based on the causal graph, provide a detailed explanation of the  
  step-by-step causal reasoning process. Finally, give the answer to the question based  
  on the causal graph and the reasoning process.",  
  "input": "Imagine a self-contained, hypothetical world with only the following
```

Ek Materyal

A Ek Metodolojik Detaylar

A.1 Cladde Veri Kümesi Örneği

Denetimli ince ayar için CausalDR veri kümelerimiz, herkese açık Cladde veri kümesine dayanmaktadır. Cladde veri kümesi, Büyük Dil Modelerinin (BDM'ler) nedensel akıl yürütme yeteneklerini değerlendirmek amacıyla başlangıçta önerilmiştir. Veri kümesi senaryolar, sorular, resmi sembolik akıl yürütme adımları ve yanıtlar sunar. Burada, veri oluşturma sürecimizle doğrudan ilgili olan özelliklerini sergiliyoruz.

Aşağıda Cladde veri kümesinden temsili bir örnek bulunmaktadır:

Orijinal Cladde Veri Kümesi Örneği

Senaryo ve Soru: Yalnızca aşağıdaki koşullara sahip, bahsedilmeyen herhangi bir faktör veya nedensel ilişki içermeyen, kendi içinde kapalı, varsayımsal bir dünya hayal edin: Koca, karısı ve çalar saat üzerinde doğrudan bir etkiye sahiptir. Karının çalar saat üzerinde doğrudan bir etkisi vardır. Çalar saat kurmayan kocalar ve çalar saat kurmayan karilar için çalan alarm olasılığı %8'dir. Çalar saat kurmayan kocalar ve çalar saat kurmayan karilar için çalan alarm olasılığı %54'tür.

Alarmı kuran kocalar ve alarmı kurmayan eşler için alarmın çalma olasılığı %41'dir. Alarmı kuran kocalar ve alarmı kurmayan eşler için alarmın çalma olasılığı %86'dır. Alarmı kurmayan kocalar için eşin alarmı kurma olasılığı %74'tür. Alarmı kuran kocalar için eşin alarmı kurma olasılığı %24'tür. Eş aracılığıyla olan aracılık etkisi göz ardı edilirse, koca çalar saatı olumlu yönde etkiler mi?

Akıl Yürütme:

X = koca; V2 = eş; Y = çalar saat olsun.

$X \rightarrow V2, X \rightarrow Y, V2 \rightarrow Y$

$$\begin{aligned} E[Y_{X=1, V2=0} - Y_{X=0, V2=0}] \\ \sum_{V2=v} P(V2 = v | X = 0) * [P(Y = 1 | X = 1, V2 = v) - P(Y = 1 | X = 0, V2 = v)] \\ P(Y = 1 | X = 0, V2 = 0) = 0.08 \\ P(Y = 1 | X = 0, V2 = 1) = 0.54 \\ P(Y = 1 | X = 1, V2 = 0) = 0.41 \\ P(Y = 1 | X = 1, V2 = 1) = 0.86 \\ P(V2 = 1 | X = 0) = 0.74 \\ P(V2 = 1 | X = 1) = 0.24 \\ 0.74 * (0.86 - 0.41) + 0.24 * (0.54 - 0.08) = 0.32 \\ 0.32 > 0 \end{aligned}$$

Cevap:Evet

Bu örnekte, senaryo değişkenler arasındaki nedensel etkileşimleri tanımlamakta, soru açıkça nedensel etkilere odaklanmakta ve akıl yürütme sembolik ve olasılıksal hesaplamalar summactadır. Bu sembolik temsiller, yapılandırılmış veri üretimimiz için nedensel düğümleri ve ilişkileri tanımlamada önemli referanslar olarak hizmet eder.

A.2 CausalDR Veri Kümesi Örneği (Bizim)

CDCR-SFT çerçevesinde denetimli ince ayar için kullanılan CausalDR veri kümesi, açık ve yapılandırılmış bir JSON formatında sunulmuştur. Her örnek üç temel bileşen içerir: (1) modelin hangi görevi gerçekleştirmesi gerektiğini açıkça yönlendiren belirli bir komut, (2) ayrıntılı senaryo açıklamasını ve spesifik nedensel akıl yürütme sorusunu sağlayan bir girdi ve (3) sunulan nedensel çizgeye açıkça dayalı, net, yapılandırılmış ve kapsamlı bir akıl yürütme sürecini gösteren bir çıktı.

Aşağıda, bu eğitim örneklerini tam olarak nasıl formüle ettiğimizi açıklamak için CausalDR veri kümesinden ayrıntılı bir örnek sunuyoruz. Her örnek yalnızca nedensel çizgeyi, çıkışım yolunu ve cevabı açıkça temsil etmeyecektir, aynı zamanda bu bileşenleri doğal ve sezgisel bir açıklayıcı anlatıya bütünsel olarak entegre eder. Bu yapılandırılmış sunum, BDM'leri nedensel akıl yürütmeye sistematik bir yaklaşım benimsemeye teşvik etmek, böylece çıkışım tutarlılığını önemli ölçüde artırmak ve halüsinsiyonları azaltmak amacıyla özel olarak tasarlanmıştır.

İnce Ayar Veri Seti Örneği (JSON)

```
{  
  "talimat": "Aşağıdaki soruya göre, senaryoyu analiz etmek için lütfen nedensel bir çizge oluşturun. Ardından, nedensel çizgeye dayanarak, adım adım nedensel akıl yürütme sürecinin ayrıntılı bir açıklamasını sunun. Son olarak, nedensel çizgeye ve akıl yürütme sürecine dayanarak sorunun cevabını verin.",  
  "girdi": "Sadece aşağıdaki koşullara sahip ve belirtilmeyen hiçbir faktör veya nedensel ilişki içermeyen, kendi kendine yeten, varsayımsal bir dünya hayal edin:
```

conditions, and without any unmentioned factors or causal relationships: Husband has a direct effect on wife and alarm clock. Wife has a direct effect on alarm clock. For husbands that don't set the alarm and wives that don't set the alarm, the probability of ringing alarm is 8%. For husbands that don't set the alarm and wives that set the alarm, the probability of ringing alarm is 54%. For husbands that set the alarm and wives that don't set the alarm, the probability of ringing alarm is 41%. For husbands that set the alarm and wives that set the alarm, the probability of ringing alarm is 86%. For husbands that don't set the alarm, the probability of alarm set by wife is 74%. For husbands that set the alarm, the probability of alarm set by wife is 24%. If we disregard the mediation effect through wife, would husband positively affect alarm clock?",

"output": "Alright, let me first review your input. Next, I will build a causal graph from the information provided, defining each node and clarifying how they interact. After that, I will detail the steps of causal inference, describing how I move from the causal graph to the final answer. To ensure clarity, I will begin by presenting the causal graph's structure, the meaning of each node, and their connections. Then, I will illustrate the inference process, leading up to the result.

Causal Graph:

First, here is the section on the causal graph nodes. For each node, I will list its ID, Name, and provide a brief description.

Nodes:

Node ID: X
 Node Name: Husband
 Node Description: Binary variable representing whether the husband sets the alarm (1=set, 0=doesn't set)
 Node ID: Y
 Node Name: Alarm Clock
 Node Description: Binary outcome variable representing whether the alarm rings (1=rings, 0=doesn't ring)
 Node ID: V2
 Node Name: Wife
 Node Description: Binary variable representing whether the wife sets the alarm (1=set, 0=doesn't set)

Next, I will explain how these nodes are linked in the causal graph, showing how information flows between them. I will go through each node in turn, indicating which nodes feed into it and which nodes it influences.

Edges:

Node: X Inputs: N/A Outputs: ['V2', 'Y']
 Node: Y Inputs: ['X', 'V2'] Outputs: N/A
 Node: V2 Inputs: ['X'] Outputs: ['Y']

Based on the current provided input information, **the inference goal** is that "Determine if Husband (X) has a positive direct effect on Alarm Clock (Y) when excluding mediation through Wife (V2)"

Given this goal, the following describes the causal inference process.

Causal Inference Process: The causal graph shows Husband (X) directly influences both Wife (V2) and Alarm (Y), while Wife (V2) also directly affects Alarm (Y). To isolate X's direct effect on Y, we calculate the natural direct effect by holding V2 constant at its X=0 distribution. Using the provided probabilities: When X=0, P(V2=1)=0.74 and P(V2=0)=0.26. The direct effect compares Y probabilities under X=1 vs X=0 for these fixed V2 values. For V2=0: 0.41 (X=1) - 0.08 (X=0) = +0.33. For V2=1: 0.86 (X=1) - 0.54 (X=0) = +0.32. Weighting these differences by V2's X=0 distribution gives (0.26*0.33) + (0.74*0.32) = 0.32. This positive result confirms Husband has a direct positive effect on Alarm Clock when excluding the Wife-mediated pathway.

As a result of this causal inference process, I will reply with the answer yes."

}

A.3 Implementation Details of DeepSeek-R1

To construct the CausalDR dataset, we selected the reasoning large language model **DeepSeek-R1** (DeepSeek-AI 2025) due to its strong reasoning capabilities and low inference cost. The model was deployed and accessed via the **DeepInfra API**¹, an inference platform designed to streamline model integration. Following recommendations from DeepSeek-R1 (DeepSeek-AI 2025), we configured the temperature at **0.6** and set the maximum token length to **8192** during dataset generation.

¹<https://deepinfra.com>

Kocanın eşi ve çalar saat üzerinde doğrudan etkisi vardır. Karının çalar saat üzerinde doğrudan bir etkisi vardır . Alarmı kurmayan kocalar ve alarmı kurmayan eşler için alarmın calma olasılığı %8'dir. Alarmı kurmayan kocalar ve alarmı kuran eşler için alarmın calma olasılığı %54'tür. Alarmı kuran kocalar ve alarmı kurmayan eşler için alarmın calma olasılığı %41'dir. Alarmı kuran kocalar ve alarmı kuran eşler için alarmın calma olasılığı %86'dır. Çalar saatı kurmayan kocalar için, eş tarafından saatin kurulma olasılığı şöyledir:

74%. Alarmı kuran kocalar için eşin alarmı kurma olasılığı %24'tür. Eş aracılığıyla olan aracılık etkisi göz ardı edilirse, koca çalar saatı olumlu yönde etkiler mi?",

"çıktı": "Pekala, öncelikle girdinizi gözden geçirmeme izin verin. Ardından, sağlanan bilgilerden bir nedensel çizge oluşturacak, her düğümü tanımlayacak ve nasıl etkileşimde bulunduklarını açıklayacağım. Daha sonra, nedensel çıkarım adımlarını detaylandırarak, nedensel çizgiden nihai cevaba nasıl ulaşımı açıklayacağım. Netliği sağlamak adına, nedensel çizgenin yapısını, her düğümün anlamını ve bağlantılarını sunarak başlayacağım. Ardından, çıkarım sürecini ve sonuca giden adımları göstereceğim.

Nedensel Çizge :

İlk olarak, nedensel çizge düğümleri bölümü aşağıdadır. Her düğüm için, Düğüm Kimliğini, Düğüm Adını ve kısa bir Düğüm Açıklamasını listeleyeceğim.

Düğümler:

Düğüm Kimliği: X
 Düğüm Adı: Koca
 Düğüm Açıklaması: Kocanın çalar saatı kurup kurmadığını temsil eden ikili değişken (1=kurar, 0=kurmaz)

Düğüm Kimliği: Y

Düğüm Adı: Çalar Saat
 Düğüm Açıklaması: Çalar saatin çalmasını (1=çalar, 0=çalmaz) temsil eden ikili sonuç değişkeni

Düğüm Kimliği: V2

Düğüm Adı: Eş
 Düğüm Açıklaması: Eşin alarmı kurup kurmamasını (1=kurar, 0=kurmaz) temsil eden ikili değişken

Ardından, bu düğümlerin nedensel çizgede nasıl bağlandığını ve aralarındaki bilgi akışını açıklayacağım . Her bir düğümü sırasıyla ele alacak, hangi düğümlerin onu etkilediğini ve hangi düğümleri etkilediğini belirteceğim.

Kenarlar:

Düğüm: X Girdiler: Yok Çıktılar: ['V2', 'Y']
 Düğüm: Y Girdiler: ['X', 'V2'] Çıktılar: Yok
 Düğüm: V2 Girdiler: ['X'] Çıktılar: ['Y']

Mevcut sağlanan girdi bilgilerine dayanarak, çıkarım hedefi "Eş (V2) aracılığıyla dolaylı etki hariç tutulduğunda , Koca'nın (X) Çalar Saat (Y) üzerinde pozitif doğrudan etkisi olup olmadığını belirleyin " şeklärindedir.

Bu hedef doğrultusunda, nedensel çıkarım süreci aşağıda açıklanmaktadır.

Nedensel Çıkarım Süreci : Nedensel çizge, Koca'nın (X) hem Eş'i (V2) hem de çalar saatı (Y) doğrudan etkilediğini, Eş'in (V2) ise çalar saatı (Y) doğrudan etkilediğini göstermektedir. X'in Y üzerindeki doğrudan etkisini ayırtmak için, V2'yi X=0 dağılımında sabit tutarak doğal doğrudan etkiyi hesaplarız. Sağlanan olasılıkları kullanarak: X=0 olduğunda, P(V2=1)=0.74 ve P(V2=0)=0.26'dır. Doğrudan etki, bu sabit V2 değerleri için X=1 ile X=0 arasındaki Y olasılıklarını karşılaştırır. V2=0 için: 0.41 (X=1) - 0.08 (X=0) = +0.33. V2=1 için: 0.86 (X =1) - 0.54 (X=0) = +0.32. Bu farkları V2'nin X=0 dağılımıyla ağırlıklandırdığımızda (0.26*0.33) + (0.74*0.32) = 0.32 elde edilir. Bu pozitif sonuç, eş aracılı yolu dışladığımızda Koca'nın çalar saat üzerinde doğrudan pozitif bir etkisi olduğunu doğrular.

Bu nedensel çıkarım süreci sonucunda , evet cevabıyla yanıt vereceğim .

}

A.3 DeepSeek-R1 Uygulama Detayları

CausalDR veri kümesini oluşturmak için, güçlü muhakeme yetenekleri ve düşük çıkarım maliyeti nedeniyle muhakeme büyük dil modeli **DeepSeek-R1** (DeepSeek-AI 2025) seçilmiştir. Model, model entegrasyonunu kolaylaştırmak için tasarlanmış bir çi- karım platformu olan **DeepInfra API**¹ aracılığıyla dağıtılmış ve erişilmiştir. DeepSeek-R1 (DeepSeek-AI 2025) tavsiyelerine uyarak, veri kümesi üretimi sırasında sıcaklığı **0.6** olarak ve maksimum jeton uzunluğunu **8192** olarak yapılandırdık.

¹<https://deepinfra.com>

A.4 Prompt Template for Generating CausalDR Data

Below, we provide the detailed prompt template used to guide the DeepSeek-R1 model in generating detailed causal graphs and explicit natural language reasoning paths for each training sample.

Detailed Prompt Template for Structured Causal Inference Data Generation

You are an expert specializing in causal inference and graph theory. Your task is to analyze a reasoning problem, construct a structured causal graph, and generate a detailed causal inference process. Your output must be in JSON format. You will receive:

- Context & Question:** A single block (`<context_question> ... </context_question>`) that contains:
 - Scenario description
 - Constraints or rules
 - Any additional details
- Reasoning:** This field contains the formal causal structure and mathematical reasoning needed to solve the problem. It includes:
 - Variable assignments (e.g., $V1 = \text{kraz}$, $X = \text{pexu}$)
 - Causal graph structure notation (e.g., $V1 \rightarrow X$, $X \rightarrow Y$)
 - Probability calculations and mathematical steps required for the solution

Your task:

- Extract causal nodes and relationships from the provided input to construct a Causal Graph.
- Causal Graph is a Directed Acyclic Graph (DAG) that represents causal influences between different variables.
- Generate a structured causal reasoning process explaining how the conclusion is derived. In the Causal Reasoning field:
 - goal:** A concise statement of the reasoning question.
 - explanation:** A paragraph that:
 - Describes variables and causal edges
 - Explains causal influence propagation
 - Translates formal math into intuition
 - Justifies the final conclusion using probabilities

Return JSON Format:

```
{  
  "Nodes": [  
    {  
      "id": "[DescriptiveVariableID]",  
      "name": "[Variable Name]",  
      "description": "[Detailed description of the causal variable]"  
    }  
    // ... more nodes if needed  
  ],  
  "Edges": [  
    {  
      "node": "[Same DescriptiveVariableID as in Nodes]",  
      "inputs": ["List of all incoming causal nodes"],  
      "outputs": ["List of all outgoing causal nodes"]  
    }  
    // ... Ensure that all Nodes are represented here.  
  ],  
  "Causal Reasoning": {  
    "goal": "[Overarching question or objective]",  
    "explanation": "[Step-by-step reasoning process]"  
  },  
  "Answer": "[yes/no]"  
}  
  
<context_question>[Insert the given scenario description, constraints, and the specific question here.</context_question>  
<reasoning>[Insert the symbolic causal graph structure, variable assignments, and mathematical reasoning steps provided by the Cladder dataset here.</reasoning>
```

A.4 CausalDR Verisi Oluşturmak için Komut İstemi Şablonu

Aşağıda, DeepSeek-R1 modelini her eğitim örneği için ayrıntılı nedensel çizgeler ve açık doğal dil muhakeme yolları oluşturması konusunda yönlendirmek amacıyla kullanılan ayrıntılı komut istemi şablonunu sunuyoruz.

Yapilandırılmış Nedensel Çıkarım Verisi Üretimi İçin Ayrıntılı Komut İstemi Şablonu

Nedensel çıkarım ve çizge teorisi konusunda uzmanlaşmış bir uzmansınız. Göreviniz, bir akıl yürütme problemini analiz etmek, yapılandırılmış bir nedensel çizge oluşturmak ve ayrıntılı bir nedensel çıkarım süreci üretmektir. Çıktınız JSON formatında olmalıdır. Alacaklarınız:

- Bağlam ve Soru:** Aşağıdakileri içeren tek bir blok (`<context question> ... </context question>`):
 - Senaryo açıklaması
 - Kısıtlamalar veya kurallar
 - Ek detaylar
- Akıl Yürütme:** Bu alan, problemi çözmek için gereken resmi nedensel yapıyı ve matematiksel akıl yürütme içeriğini kapsar. Sunları kapsar:
 - Değişken atamaları (örn. $V1 = \text{kraz}$, $X = \text{pexu}$)
 - Nedensel çizge yapı gösterimi (örn. $V1 \rightarrow X$, $X \rightarrow Y$)
 - Çözüm için gerekli olasılık hesaplamaları ve matematiksel adımlar

Göreviniz:

- Sağlanan girdiden nedensel düğümleri ilişkileri çıkararak bir Nedensel Çizge oluşturun.
- Nedensel Çizge, farklı değişkenler arasındaki nedensel etkileşimleri temsil eden Yönlendirilmiş Döngüsel Çizgedir (DAG).
- Sonucun nasıl elde edildiğini açıklayan yapılandırılmış bir nedensel akıl yürütme süreci üretin. Nedensel Akıl Yürütme alanında:
 - hedef:** Akıl yürütme sorusunun özü bir ifadesi.
 - açıklama:** Aşağıdakileri içeren bir paragraf:
 - Değişkenleri ve nedensel kenarları tanımlar
 - Nedensel etki yayılmasını açıklar
 - Resmi matematiği sezgisel bir dille aktarır
 - Nihai sonucu olasılıkları kullanarak gerekçelendirir

JSON Formatında Dönüş:

```
{  
  "Nodes": [  
    {  
      "id": "[TanımlayıcıDeğişkenKimliği]",  
      "name": "[Değişken Adı]",  
      "description": "[Nedensel değişkenin ayrıntılı açıklaması]"  
    }  
    // ... gerekirse daha fazla düğüm  
  ],  
  "Edges": [  
    {  
      "node": "[Düğümlerdekiyle aynı TanımlayıcıDeğişkenKimliği]",  
      "inputs": ["Tüm gelen nedensel düğümlerin listesi"],  
      "outputs": ["Tüm giden nedensel düğümlerin listesi"]  
    }  
    // ... Tüm düğümlerin burada temsil edildiğinden emin olun.  
  ],  
  "Causal Reasoning": {  
    "goal": "[Genel soru veya amaç]",  
    "explanation": "[Adım adım akıl yürütme süreci]"  
  },  
  "Answer": "[evet/hayır]"  
}  
  
<bağlam sorusu> [Verilen senaryo açıklamasını, kısıtlamaları ve spesifik soruyu buraya ekleyin. </bağlam sorusu> <reasoning> [Cladder veri kümesi tarafından sağlanan sembolik nedensel çizge yapısını, değişken atamalarını ve matematiksel akıl yürütme adımlarını buraya ekleyin.] </reasoning>
```

A.5 Detailed Algorithm for Fine-tuning Dataset Construction

We provide a detailed pseudocode representation of our automated dataset construction pipeline in Algorithm 1. This algorithm clearly illustrates the structured process of generating high-quality fine-tuning data leveraging the DeepSeek-R1 model. Each step explicitly ensures data correctness, coherence, and suitability for supervised fine-tuning within our proposed framework.

Algorithm 1: CausalDR Dataset Construction

```

1: Input: Cladder training set  $D_{\text{Cladder}} = \{(c_i, q_i, r_i, a_i)\}_{i=1}^N$ , where:
    $c_i$ : scenario context;
    $q_i$ : causal inference question;
    $r_i$ : symbolic reasoning from Cladder dataset;
    $a_i$ : ground-truth answer.
   LLM for causal reasoning: DeepSeek-R1; maximum attempts  $K = 15$ 
2: Output: Fine-tuning dataset  $D_{\text{CausalDR}}$ 
3: Initialize  $D_{\text{CausalDR}} \leftarrow \emptyset$ 
4: for all  $(c_i, q_i, r_i, a_i) \in D_{\text{Cladder}}$  do
5:   Construct structured prompt  $p_i$  from  $(c_i, q_i, r_i)$  (details in Appx. A.4)
6:   Set success  $\leftarrow$  False,  $k \leftarrow 0$ 
7:   while  $\neg$ success  $\wedge k < K$  do
8:      $(G_i, P_i, y_i) \leftarrow \text{DeepSeek-R1}(p_i)$                                  $\triangleright$  structured JSON output
9:     if  $y_i = a_i$  then
10:    Construct coherent inference paragraph  $S_i$  by explicitly integrating  $(G_i, P_i, y_i)$  into a natural explanatory narrative (See example output in Appx. A.2).
11:     $D_{\text{CausalDR}} \leftarrow D_{\text{CausalDR}} \cup \{(c_i, q_i, S_i)\}$                  $\triangleright S_i$  encapsulates  $G_i, P_i, y_i$ 
12:    success  $\leftarrow$  True
13:   end if
14:    $k \leftarrow k + 1$ 
15: end while
16: end for
17: return  $D_{\text{CausalDR}}$ 

```

Algorithm 1 systematically describes how we integrate DeepSeek-R1 to produce a coherent reasoning sequence (S) comprising structured causal graphs (G), explicit inference paths (P), and answers (y), thereby ensuring the resulting fine-tuning dataset addresses the challenge of hallucinations in LLM inference.

B Additional Experimental Details

B.1 Baselines Methods

To assess the effectiveness of our CDCR-SFT method, we compare it against 5 commonly used reasoning methods. While these methods have achieved some success in reasoning, they still have limitations in dealing with complex causal reasoning tasks, including the difficulty of effectively capturing causal relationships, as described in the Introduction section. Specifically, our comparative approach consists of:

- **Chain-of-Thought (CoT)** (Wei et al. 2022): CoT instructs the model to generate intermediate reasoning steps, helping models to solve complex problems by decomposing tasks into simpler sub-steps. We use zero-shot CoT without any few-shot prompting (Kojima et al. 2023), only the reasoning prompt, and refer to it as CoT in the following sections.
- **Chain-of-Thought Self-Consistency (CoT-SC)** (Wang et al. 2023): An improved version of CoT that samples multiple reasoning paths and selects the final answer based on consistency among these paths.
- **Causal Chain-of-Thought (CausalCoT)** (Jin et al. 2023): CausalCoT guides the model through defined steps, including causal graph extraction, formalization of queries, and calculation of counterfactual outcomes.
- **Tree-of-Thoughts (ToT)** (Yao et al. 2023): ToT organizes the reasoning process as a tree structure, enabling the model to explore multiple reasoning paths.
- **Graph-of-Thoughts (GoT)** (Besta et al. 2024): GoT organizes reasoning steps into a graph structure, modeling each reasoning step as a node and dependencies among these steps as edges, without explicitly modeling causal relationships.

B.2 Detailed Validation of Cladder Test Set

To ensure the quality and validity of our Cladder test set, we implemented a rigorous, two-step validation process involving both automated evaluation with DeepSeek-R1 and manual verification. Here, we describe this validation workflow in detail.

A.5 İnce Ayar Veri Kümesi Oluşturma için Detaylı Algoritma

Otomatik veri kümesi oluşturma arduşik düzennimizin detaylı bir sözde kod gösterimini Algoritma 1'de sunuyoruz. Bu algoritma, DeepSeek-R1 modelini kullanarak yüksek kaliteli ince ayar verisi üretme sürecini açıkça göstermektedir. Her adım, önerilen çerçeveyimiz dahilinde verilerin doğruluğunu, tutarlılığını ve denetimli ince ayar için uygunluğunu açıkça sağlamaktadır.

Algoritma 1: CausalDR Veri Kümesi Oluşturma

```

1: Girdi: Cladder eğitim seti  $D_{\text{Cladder}} = \{(c_i, q_i, r_i, a_i)\}_{i=1}^N$ , burada:
    $c_i$ : senaryo bağlantı;
    $q_i$ : nedensel çıkarım sorusu;
    $r_i$ : Cladder veri setinden sembolik akıl yürütme;
    $a_i$ : zemin gerceği yanıtı.
   Nedensel akıl yürütme için BDM: DeepSeek-R1; maksimum deneme sayısı  $K = 15$ 
2: Çıktı: İnce ayar veri seti  $D_{\text{CausalDR}}$ 
3: Başlat  $D_{\text{CausalDR}} \leftarrow \emptyset$ 
4:  $(c_i, q_i, r_i, a_i) \in D_{\text{Cladder}}$  do tümü için
5: Yapılandırılmış komut istemi oluştur  $p_i(c_i, q_i, r_i)$ 'dan (ayrintılar Ek A.4'te)
6: Başarıyı ayarla  $\leftarrow$  Yanlış,  $k \leftarrow 0$ 
7: iken  $\neg$ başarılı  $\wedge k < K$  do
8:  $(G_i, P_i, y_i) \leftarrow \text{DeepSeek-R1}(p_i)$                                  $\triangleright$  yapılandırılmış JSON çıktı
9: eğer  $y_i = a_i$  ise o zaman
10: Tutarlı çıkarım paragrafi  $S_i$  oluşturun,  $(G_i, P_i, y_i)$  bilgilerini açıkça doğal bir açıklayıcı anlatıma entegre ederek (Ek A.2.'deki örnek çıktıya bakınız).
11:  $D_{\text{CausalDR}} \leftarrow D_{\text{CausalDR}} \cup \{(c_i, q_i, S_i)\}$                  $\triangleright S_i$  şunları kapsar:  $G_i, P_i, y_i$ 
12: başarı  $\leftarrow$  Doğru
13: eğer sonu
14:  $k \leftarrow k + 1$ 
15: döngü sonu
16: döngü sonu
17:  $D_{\text{CausalDR}}$ döndür

```

Algoritma 1, DeepSeek-R1'i yapılandırılmış nedensel çizgeler (G), açık çıkarım yolları (P) ve yanıtlar (y) içeren tutarlı bir akıl yürütme dizisi (S) üretmek için nasıl entegre ettiğimizi sistematik olarak açıklayarak, ortaya çıkan ince ayar veri setinin BDM çıkarımındaki halüsinasyon sorununu ele almasını sağlar.

B Ek Deneysel Detaylar

B.1 Temel Yöntemler

CDCR-SFT yöntemimizin etkinliğini değerlendirmek için, onu yaygın olarak kullanılan 5 akıl yürütme yöntemiyle karşılaştırıyoruz. Bu yöntemler akıl yürütmede bazı başarılar elde etmiş olsa da, Giriş bölümünde açıkladığı gibi, nedensel ilişkileri etkili bir şekilde yakaladıktan zorluk da dâhil olmak üzere karmaşık nedensel akıl yürütme görevleyle başa çıkmada hâlâ sınırlamalara sahiptir. Özellikle, karşılaşmamızın şunlardan oluşmaktadır:

- **Düşünce Zinciri (CoT)** (Wei vd. 2022): Düşünce Zinciri, modeli ara akıl yürütme adımları oluşturmaya yönlendirerek, görevleri daha basit alt adımlara ayırarak karmaşık problemleri çözmeye yardımcı olur. Hiçbir örneklem öncülüğu olmadan sıfır-shot Düşünce Zinciri (Kojima vd. 2023) kullanıyoruz, yalnızca akıl yürütme öncülüünü kullanıyoruz ve sonraki bölgelerde buna Düşünce Zinciri olarak atıfta bulunuyoruz.
- **Düşünce Zinciri Kendini Tutarlılığı (CoT-SC)** (Wang vd. 2023): Birden çok akıl yürütme yolu örnekleyle ve bu yollar arasındaki tutarlılığı dayanarak nihai cevabı seçen Düşünce Zinciri'nin geliştirilmiş bir versiyonu.
- **Nedensel Düşünce Zinciri (CausalCoT)** (Jin vd. 2023): CausalCoT, nedensel çizge çıkarımı, sorguların biçimlendirilmesi ve karşıolgusal sonuçların hesaplanması dahil olmak üzere tanımlanmış adımlar aracılığıyla modeli yönlendirir.
- **Düşünce Ağacı (ToT)** (Yao vd. 2023): ToT, akıl yürütme sürecini bir ağaç yapısı olarak düzenleyerek modelin birden çok akıl yürütme yolunu keşfetmesini sağlar.
- **Düşünce Çizgesi (GoT)** (Besta ve diğerleri. 2024): GoT, akıl yürütme adımlarını bir çizge yapısı halinde düzenler; her bir akıl yürütme adımı bir düğüm, bu adımlar arasındaki bağımlılıkları ise kenarlar olarak modeller, ancak nedensel ilişkileri açıkça modellemez.

B.2 Cladder Test Setinin Ayrıntılı Doğrulaması

Cladder test setimizin kalitesini ve geçerliliğini sağlamak amacıyla DeepSeek-R1 ile otomatik değerlendirme ve manuel doğrulama olmak üzere titiz, iki aşamalı bir doğrulama süreci uyguladık. Bu doğrulama iş akışını burada ayrıntılı olarak açıklıyoruz.

Step 1: Automated Validation via DeepSeek-R1. We constructed prompts that asked the DeepSeek-R1 model to verify the correctness of the provided reasoning and answer for each test sample. An example validation prompt is as follows:

Example Validation Prompt

```
You are an expert analyzing causal reasoning. Evaluate if the reasoning process and answer are correct for this causal inference problem.  
<context_question>...[Scenario and causal question here]...</context_question>  
<reasoning>...[Provided symbolic reasoning steps]...</reasoning>  
<proposed_answer>...[Provided answer]...</proposed_answer>  
Provide a JSON response:  
{  
    "reasoning_valid": true/false,  
    "reasoning_error": "Brief description of error if any, otherwise 'None'",  
    "answer_correct": true/false,  
    "correct_answer": "yes/no",  
    "brief_explanation": "1-2 sentences explaining your assessment"  
}
```

If either the reasoning or the answer was marked incorrect, these samples were flagged for further review.

Step 2: Manual Verification. Samples flagged as problematic by DeepSeek-R1 underwent manual review by domain experts to confirm the validity of the model’s assessment. During this review, we carefully inspected reasoning accuracy and answer correctness, retaining only those samples unanimously confirmed as valid and logically sound.

Through this rigorous validation pipeline, we removed a total of 189 problematic samples, refining our test set down to 1,906 high-quality examples, suitable for rigorous causal reasoning evaluation.

The final validated Cladder test set, along with the validation scripts, will be publicly released to ensure reproducibility of our experiments.

B.3 WIQA Question Reformulation Procedure

Many original WIQA questions contained ambiguous phrasing or grammatical errors, potentially affecting evaluation results. An example of such ambiguity is:

Original Question (Ambiguous Example):

“Suppose less DNA available happens, how will it affect hurting the DNA to replicate properly?”
Options: A) more, B) less, C) no effect

To eliminate such issues, we reformulated all questions into a clear, standardized format, strictly matching the provided answer options. The improved version of the above question is:

Improved Question (Reformulated Example):

“Will having less available DNA cause more replication errors, fewer replication errors, or have no effect?”
Options: A) more, B) less, C) no effect

We used an automated approach employing the DeepSeek-R1 to rewrite each selected WIQA question. Below is the prompt we employed for the automatic rewriting:

Adım 1: DeepSeek-R1 Aracılığıyla Otomatik Doğrulama. DeepSeek-R1 modelinden, her test örneği için sağlanan akıl yürütme ve cevabın doğruluğunu teyit etmesini isteyen istemler oluşturduk. Bir örnek doğrulama istemi aşağıdaki gibidir:

Örnek Doğrulama İstemci

```
Nedensel akıl yürütmemeyi analiz eden bir uzmansınız. Bu nedensel çıkarım problemi için akıl yürütme sürecinin ve cevabin doğru olup olmadığını değerlendirin.  
<context_question> ...[Senaryo ve nedensel soru burada]...</context_question>  
<reasoning> ...[Sağlanan象征ik akıl yürütme adımları]...</reasoning>  
<proposed_answer> ...[Sağlanan cevap]...</proposed_answer>  
Bir JSON yanıtı sağlayın:  
{  
    "reasoning_valid": true/false,  
    "reasoning_error": "Varsa hatanın kısa açıklaması, aksi takdirde 'Yok'", "answer_correct": true/false  
    "correct_answer": "evet/hayır",  
    "brief_explanation": "Değerlendirmenizi açıklayan 1-2 cümle"  
}
```

Akıl yürütme veya cevaptan herhangi biri yanlış olarak işaretlenirse, bu örnekler daha fazla inceleme için işaretlendi.

Adım 2: Manuel Doğrulama. DeepSeek-R1 tarafından sorunlu olarak işaretlenen örnekler, model değerlendirmesinin doğrulanmak amacıyla alan uzmanları tarafından manuel olarak incelenmiştir. Bu inceleme sırasında, muhakeme doğruluğunu ve yanıtın kesinliğini dikkatle denetleyerek, yalnızca oybirliğiyle geçerli ve mantıksal olarak sağlam olduğu teyit edilen örnekleri muhafaza ettik.

Bu titiz doğrulama süreciyle toplam 189 sorunlu örneği çıkararak, test setimizi titiz nedensel akıl yürütme değerlendirmesi için uygun olan 1.906 yüksek kaliteli örneğe indirdigedik.

Nihai doğrulanmış Cladder test seti, doğrulama betikleriyle birlikte deneylerimizin tekrarlanabilirliğini sağlamak amacıyla herkese açık olarak yayımlanacaktır.

B.3 WIQA Soru Yeniden Biçimlendirme Prosedürü

Birçok orijinal WIQA sorusu, değerlendirme sonuçlarını potansiyel olarak etkileyen belirsiz ifadeler veya dilbilgisi hataları içermektedir. Böyle bir belirsizlik örneği şu şekildedir:

Orijinal Soru (Belirsiz Örnek):

“Varsayılmı ki daha az DNA mevcut olduğunda, bu durum DNA'nın düzgün bir şekilde çoğalmasını nasıl etkiler?”
Seçenekler: A) daha fazla, B) daha az, C) etkisi yok

Bu tür sorunları gidermek için tüm soruları, verilen yanıt seçenekleriyle birebir eşleşen, açık ve standartlaştırılmış bir formata yeniden düzenlenledik. Yukarıdaki sorunun geliştirilmiş versiyonu şu şekildedir:

Geliştirilmiş Soru (Yeniden Formüle Edilmiş Örnek):

“Daha az DNA'ya sahip olmak daha fazla replikasyon hatasına mı, daha az replikasyon hatasına mı yol açar, yoksa hiçbir etkisi olmaz mı?”
Seçenekler: A) daha fazla, B) daha az, C) etkisi yok

Seçilen her WIQA sorusunu yeniden yazmak için DeepSeek-R1 kullanan otomatik bir yaklaşım uyguladık. Otomatik yeniden yazma için kullandığımız istem aşağıdadır:

Prompt Used for Question Reformulation

I have an English multiple-choice question with incorrect grammar and unclear meaning. I know that the correct answer is "Answer choice". Please help me rewrite this question so that:

- It is grammatically correct.
- It is logically clear and specific.
- It introduces no new information from outside the paragraph.
- It strictly preserves the original multiple-choice options:

A) more, B) less, C) no effect.

The question must be rewritten exactly in the following format:

"Will [cause/change] cause more [effect], fewer [effect], or have no effect?"

This format must match the options exactly and avoid ambiguity. The rewritten question should clearly express the potential impact of a change on a specific outcome, and the options "more", "less", and "no effect" should directly correspond to the parts of the question.

Here is the background context:

Process steps:

{List of paragraph steps provided here}

Original question:

{Original problematic question provided here}

Options:

- A) more
- B) less
- C) no effect

Correct answer: {Correct answer choice provided here}

Return your result strictly in the following JSON format:

```
{  
  "improved_question": "Your improved question here"  
}
```

All 212 selected WIQA questions (106 from INPARA_EFFECT and 106 from EXOGENOUS_EFFECT) underwent this reformulation. A random subset of reformulated questions was manually reviewed to confirm grammatical correctness and logical clarity.

B.4 Details of LLM-based Evaluation for Causal Graph Quality

We employed an automatic evaluation approach utilizing an LLM (gpt-4o-mini) as a judge to assess the quality of the generated causal graphs. Specifically, given a causal reasoning context, a ground-truth causal graph, and a model-generated causal graph, the evaluator rated each generated graph along three dimensions: **Node Accuracy**, **Edge Accuracy**, and **Structural Fidelity**, assigning scores on a scale from 0 to 10. To ensure reproducibility, we set the inference temperature of gpt-4o-mini to 0.

Causal Graph Quality Scoring Criteria The detailed scoring criteria for these dimensions are presented in Table 5.

Prompt Template for LLM-based Evaluation The prompt template used for this automatic evaluation was as follows:

Prompt Template for LLM-based Evaluation

You are an expert evaluator specialized in assessing the quality of causal graph structures.

Your task:

Given a specific causal reasoning scenario (the problem context), along with a Ground Truth causal graph description (serving as the evaluation standard), your goal is to evaluate the quality of a **model-generated causal graph** (the evaluation target).

Soru Yeniden Formülasyonu İçin Kullanılan İstem

Yanlış dil bilgisine ve belirsiz anlamaya sahip bir İngilizce çoktan seçmeli sorum var. Doğru cevabin " { Cevap seçenekleri } " olduğunu biliyorum. Lütfen bu soruyu aşağıdaki şekilde yeniden yazmama yardım edin:

- Dil bilgisi açısından doğru olsun.
- Mantıksal olarak açık ve spesifik olsun.
- Paragraf dışından yeni bilgi içermesin.
- Orijinal çoktan seçmeli seçenekleri kesinlikle korusun:

A) daha fazla, B) daha az, C) hiçbir etkisi yok.

- Soru kesinlikle aşağıdaki formatta yeniden yazılmalıdır:

"[neden/değişiklik] daha fazla [etki]ye mi, daha az [etki]ye mi neden olacak, yoksa hiçbir etkisi olmayacak mı?"

Bu format, seçeneklerle tam olarak eşleşmeli ve belirsizlikten kaçınmalıdır. Yeniden yazılan soru, bir değişikliğin belirli bir sonuç üzerindeki potansiyel etkisini açıkça ifade etmeli ve "daha fazla", "daha az" ve "bir etki" seçenekleri, sorunun ilgili kısımlarına doğrudan karşılık gelmelidir.

İşte arka plan bağlamı:

İşlem adımları:

{Paragraf adımlarının listesi burada verilmiştir}

Orijinal soru:

{Orijinal sorunlu soru burada verilmiştir}

Seçenekler:

- A) daha fazla
- B) daha az
- C) bir etki

Doğru cevap: {Doğru cevap seçeneği burada verilmiştir}

Sonucunuza kesinlikle aşağıdaki JSON formatında döndürün:

```
{  
  "iyileştirilmiş_soru": "İyileştirilmiş sorunuz burada"  
}
```

Seçilen tüm 212 WIQA sorusu (106'sı INPARA EFFECT ve 106'sı EKSOJEN EFFECT'ten) bu yeniden formülasyon sürecinden geçmiştir. Yeniden formüle edilmiş soruların rastgele bir alt kümesi, dilbilgisel doğruluğu ve mantıksal netliği teyit etmek amacıyla manuel olarak incelenmiştir.

B.4 Nedensel Çizge Kalitesi için BDM Tabanlı Değerlendirmenin Detayları

Üretilen nedensel çizgelerin kalitesini değerlendirmek amacıyla bir BDM'yi (gpt-4o-mini) bir değerlendirme aracı olarak kullanan otomatik bir değerlendirme yaklaşımı uyguladık. Özellikle, belirli bir nedensel akıl yürütme bağlamı, bir zemin gerceği nedensel çizgesi ve model tarafından üretilen bir nedensel çizge verildiğinde, değerlendirme aracı her üretilen çizgeyi üç boyutta (Düğüm Doğruluğu, Kenar Doğruluğu ve Yapısal Doğruluk) 0'dan 10'a kadar bir olçeğe puanlayarak değerlendirmiştir. Tekrarlanabilirliği sağlamak amacıyla, gpt-4o-mini'nin çıkışım sıklığını 0 olarak ayarladık.

Nedensel Çizge Kalitesi Puanlama Kriterleri Bu boyutlara ilişkin detaylı puanlama kriterleri Tablo 5'te sunulmaktadır.

BDM Tabanlı Değerlendirme İçin İstek Şablonu Bu otomatik değerlendirme için kullanılan istek şablonu aşağıdaki gibidir:

BDM Tabanlı Değerlendirme İçin İstek Şablonu

Siz, nedensel çizge yapılarının kalitesini değerlendirme konusunda uzmanlaşmış bir değerlendirme uzmanınızı.

Göreviniz:

Belirli bir nedensel akıl yürütme senaryosu (problem bağlamı) ve bir Zemin Gerceği nedensel çizge açıklaması (değerlendirme standartı olarak hizmet veren) verildiğinde, hedefiniz model tarafından oluşturulan bir nedensel çizgenin (değerlendirme hedefi) kalitesini değerlendirmektir.

Table 5: Detailed scoring criteria for LLM-based causal graph quality evaluation.

Score	Node Accuracy	Edge Accuracy	Structural Fidelity
10	All nodes perfectly identified, no errors or omissions.	All edges (including directions) identified perfectly.	Structure perfectly matches the Ground Truth; fully reasonable.
9	All core nodes correctly identified; only minor discrepancies with non-critical nodes.	Nearly perfect; only one minor discrepancy on a non-critical edge.	Structure highly matches, minor irrelevant differences only.
8	Nearly all nodes correctly identified; only 1 minor node omitted or misidentified.	Nearly all edges correct; just 1 minor edge omitted or misidentified.	Structure largely matches; minor differences but no significant flaws.
7	Core nodes identified accurately, but minor omissions or misidentifications (1–2 non-critical nodes).	Most core edges correct; 1–2 non-critical edges missed or incorrect.	Clearly reasonable and coherent structure, minor noticeable flaws.
6	Most nodes correct, but clearly missing or misidentifying a few nodes.	Generally correct, but clearly missing or incorrectly identifying 1–2 important edges.	Generally reasonable but with clear structural errors or omissions.
5	Around half of the nodes correct; obvious omissions or errors.	Around half of the edges correct; obvious errors or omissions.	Obvious structural problems; overall logic still somewhat coherent.
4	Only a small portion of nodes correct; many omissions or errors.	Poorly identified; only a small portion of core edges correct.	Partially confusing; only some parts clearly reasonable.
3	Most nodes incorrect, only a few correct.	Mostly incorrect edges, only a few correct.	Mostly chaotic; few structurally reasonable elements.
2	Mostly incorrect; only one or two nodes correct by chance.	Only 1–2 edges correct.	Severe structural issues; only minor elements reasonable by chance.
1	Only one node identified correctly; all others wrong.	Almost entirely incorrect; only one edge correct by chance.	Nearly completely incorrect; minimal structural coherence by chance.
0	Completely incorrect; no correct nodes identified.	Completely incorrect; no correct edges identified.	Completely incorrect; no structural coherence at all.

Tablo 5: BDM tabanlı nedensel çizge kalite değerlendirmesi için detaylı puanlama kriterleri.

Puan	Düğüm Doğruluğu	Kenar Doğruluğu	Yapısal Uygunluk
10	Tüm düğümler hatalı veya eksiksiz olmaksızın mükemmel şekilde tanımlanmıştır.	Tüm kenarlar (yönler dahil) mükemmel şekilde tanımlanmıştır.	Yapı Zemin Gerçeği ile mükemmel uyumlu; tamamen makul.
9	Tüm ana düğümler doğru bir şekilde belirlendi; yalnızca kritik olmayan düğümlerde küçük bir küçük tutarsızlıklar var.	Neredeyse mükemmel; kritik olmayan bir kenarda sadece bir küçük tutarsızlık var.	Yapı büyük ölçüde uyumlu, yalnızca öünsüz, alakasız farklılıklar mevcut.
8	Neredeyse tüm düğümler doğru bir şekilde belirlendi; yalnızca 1 küçük düğüm eksik veya yanlış tanımlanmış.	Neredeyse tüm kenarlar doğru; sadece 1 ikincil kenar atlanmış veya yanlış tanımlanmış.	Yapı büyük ölçüde eşleşiyor; küçük farklılıklar mevcut ancak önemli bir kusur bulunmuyor.
7	Temel düğümler doğru bir şekilde belirlenmiş, ancak küçük eksikslikler veya yanlış tanımlamalar (1-2 kritik olmayan düğüm) mevcut.	Çoğu temel kenar doğru; 1-2 kritik olmayan kenar eksik veya hatalı.	Açıkça makul ve tutarlı bir yapı, ancak küçük fark edilebilir kusurlar var.
6	Çoğu düğüm doğru, ancak birkaç düğüm açıkça eksik veya yanlış tanımlanmış.	Genel olarak doğru, ancak 1-2 önemli kenar açıkça eksik veya yanlış belirlenmiş.	Genel olarak makul ancak açık yapısal hatalar veya eksikslikler içeriyor.
5	Düğümlerin yaklaşık yarısı doğru; bariz eksikslikler veya hatalar.	Kenarların yaklaşık yarısı doğru; bariz hatalar veya eksikslikler.	Bariz yapısal sorunlar; genel mantık hâlâ bir ölçüde tutarlı.
4	Düğümlerin sadece küçük bir kısmı doğru; birçok eksikslik veya hata.	Kötü tanımlanmış; çekirdek kenarların sadece küçük bir kısmı doğru.	Kısmen kafa karıştırıcı; sadece bazı kısımları açıkça makul.
3	Çoğu düğüm yanlış, sadece birkaç doğru.	Çoğunlukla yanlış kenarlar, sadece birkaç doğru.	Çoğunlukla kaotik; birkaç yapısal olarak makul öğe.
2	Çoğunlukla yanlış; şansa bağlı olarak sadece bir veya iki düğüm doğru.	Yalnızca 1-2 kenar doğru.	Ciddi yapısal sorunlar; sadece küçük elementler şans eseri makul.
1	Sadece bir düğüm doğru tanımlanmış; diğerleri yanlış.	Neredeyse tamamen yanlış; şansa bağlı olarak sadece bir kenar doğru.	Neredeyse tamamen yanlış; şans eseri minimum yapısal tutarlılık.
0	Tamamen yanlış; doğru düğüm belirlenmedi.	Tamamen yanlış; doğru kenar belirlenmedi.	Tamamen yanlış; hiçbir yapısal tutarlılık yok.

Important Clarifications:

- You are to assign scores specifically to the model-generated causal graph, NOT to the Ground Truth causal graph.
- Your evaluation must be strictly based on comparing the model-generated causal graph with the provided Ground Truth causal graph and guided by the causal reasoning problem context, which clarifies the meaning of each node and edge.
- Evaluate separately along three independent dimensions:
 - Node Accuracy(0–10 points)
 - Edge Accuracy(0–10 points)
 - Overall Structural Quality(0–10 points)
- Follow the detailed scoring criteria provided below, and briefly justify your rating for each dimension.

Detailed Scoring Criteria (0–10 points each dimension):

Node Accuracy:

- 10: All nodes perfectly identified, no errors or omissions.
- 9: All core nodes correctly identified; only minor discrepancies with non-critical nodes.
- 8: Nearly all nodes correctly identified; only 1 minor node omitted or misidentified.
- 7: Core nodes identified accurately, but minor omissions or misidentifications (1–2 non-critical nodes).
- 6: Most nodes correct, but clearly missing or misidentifying a few nodes.
- 5: Around half of the nodes correct; obvious omissions or errors.
- 4: Only a small portion of nodes correct; many omissions or errors.
- 3: Most nodes incorrect, only a few correct.
- 2: Mostly incorrect; only one or two nodes correct by chance.
- 1: Only one node identified correctly; all others wrong.
- 0: Completely incorrect; no correct nodes identified.

Edge Accuracy:

- 10: All edges (including directions) identified perfectly.
- 9: Nearly perfect; only one minor discrepancy on a non-critical edge.
- 8: Nearly all edges correct; just 1 minor edge omitted or misidentified.
- 7: Most core edges correct; 1–2 non-critical edges missed or incorrect.
- 6: Generally correct, but clearly missing or incorrectly identifying 1–2 important edges.
- 5: Around half of the edges correct; obvious errors or omissions.
- 4: Poorly identified; only a small portion of core edges correct.
- 3: Mostly incorrect edges, only a few correct.
- 2: Only 1–2 edges correct.
- 1: Almost entirely incorrect; only one edge correct by chance.
- 0: Completely incorrect; no correct edges identified.

Overall Structural Quality:

- 10: Structure perfectly matches the Ground Truth; fully reasonable.
- 9: Structure highly matches; minor irrelevant differences only.
- 8: Structure largely matches; minor differences but no significant flaws.
- 7: Clearly reasonable and coherent structure, minor noticeable flaws.
- 6: Generally reasonable but with clear structural errors or omissions.
- 5: Obvious structural problems; overall logic still somewhat coherent.
- 4: Partially confusing; only some parts clearly reasonable.
- 3: Mostly chaotic; few structurally reasonable elements.
- 2: Severe structural issues; only minor elements reasonable by chance.
- 1: Nearly completely incorrect; minimal structural coherence by chance.
- 0: Completely incorrect; no structural coherence at all.

Please strictly follow the JSON format below when returning your evaluation:

```
{  
  "Node_Accuracy": {"Score": (0–10), "Brief_Reasoning": "..."},  
  "Edge_Accuracy": {"Score": (0–10), "Brief_Reasoning": "..."},  
  "Overall_Structural_Quality": {"Score": (0–10), "Brief_Reasoning": "..."}  
}
```

Now, proceed to your evaluation:

Önemli Açıklamalar:

- Puanları özellikle model tarafından oluşturulan nedensel çizgeye atamalısınız, Zemin Gerçeği nedensel çizgesine DEĞİL.
- Değerlendirmeniz, model tarafından oluşturulan nedensel çizgenin sağlanan Zemin Gerçeği nedensel çizgesiyle karşılaştırılmasına kesinlikle dayanmalı ve her düğüm ile kenarın anlamını açıklayan nedensel akıl yürütme problem bağlamı tarafından yönlendirilmelidir.
- Üç bağımsız boyutta ayrı ayrı değerlendirin:
- Düğüm Doğruluğu (0–10 puan)
- Kenar Doğruluğu (0–10 puan)
- Genel Yapısal Kalite (0–10 puan)

- Aşağıda verilen ayrıntılı puanlama kriterlerini takip edin ve her boyut için derecelendirmenizi kısaca gerekçelendirin. Ayrıntılı Puanlama Kriterleri (her boyut 0–10 puan):

Düğüm Doğruluğu:

- 10: Tüm düğümler kusursuzca belirlendi, hata veya eksik yok.
- 9: Tüm ana düğümler doğru bir şekilde belirlendi; yalnızca kritik olmayan düğümlerle ilgili küçük tutarsızlıklar.
- 8: Neredeyse tüm düğümler doğru şekilde tanımlanmıştır; yalnızca 1 küçük düğüm atlanmış veya yanlış tanımlanmıştır.
- 7: Çekirdek düğümler doğru tanımlanmış, ancak küçük atlamalar veya yanlış tanımlamalar (1–2 kritik olmayan düğüm) mevcuttur.
- 6: Çoğu düğüm doğru, ancak birkaç düğüm açıkça eksik veya yanlış tanımlanmıştır.
- 5: Düğümlerin yaklaşık yarısı doğrudur; bariz atlamalar veya hatalar vardır.
- 4: Düğümlerin yalnızca küçük bir kısmı doğrudur; birçok atlama veya hata vardır.
- 3: Çoğu düğüm yanlış, sadece birkaç doğrudur.
- 2: Çoğunlukla yanlış; yalnızca bir veya iki düğüm şans eseri doğrudur.
- 1: Yalnızca bir düğüm doğru tanımlanmıştır; diğerleri yanlışdır.
- 0: Tamamen yanlış; hiçbir doğru düğüm tanımlanmamıştır.

Kenar Doğruluğu:

- 10: Tüm kenarlar (yönerler dahil) mükemmel şekilde tanımlanmıştır.
- 9: Neredeyse mükemmel; kritik olmayan bir kenarda yalnızca küçük bir tutarsızlık vardır.
- 8: Neredeyse tüm kenarlar doğru; yalnızca 1 küçük kenar atlanmış veya yanlış tanımlanmıştır.
- 7: Çoğu temel kenar doğru; 1–2 kritik olmayan kenar atlanmış veya yanlış.
- 6: Genellikle doğru, ancak 1–2 önemli kenar açıkça eksik veya yanlış tanımlanmış.
- 5: Kenarların yaklaşık yarısı doğru; bariz hatalar veya eksiklikler mevcut.
- 4: Kötü tanımlanmış; temel kenarların yalnızca küçük bir kısmı doğru.
- 3: Çoğu kenar yanlış, yalnızca birkaç doğru.
- 2: Yalnızca 1–2 kenar doğru.
- 1: Neredeyse tamamen yanlış; yalnızca bir kenar şans eseri doğru.
- 0: Tamamen yanlış; hiçbir doğru kenar tanımlanmamıştır.

Genel Yapısal Kalite:

- 10: Yapı Zemin Gerçeği ile mükemmel şekilde eşleşiyor; tamamen makul.
- 9: Yapı yüksek düzeyde eşleşiyor, yalnızca küçük alakasız farklılıklar var.
- 8: Yapı büyük ölçüde eşleşiyor; küçük farklılıklar var ancak önemli kusurlar yok.
- 7: Açıkça makul ve tutarlı yapı, fark edilebilir küçük kusurlar.
- 6: Genellikle makul ancak belirgin yapısal hatalar veya eksiklikler içeriyor.
- 5: Bariz yapısal sorunlar; genel mantık hala bir dereceye kadar tutarlı.
- 4: Kismen kafa karıştırıcı; sadece bazı kısımlar açıkça makul.
- 3: Çoğunlukla kaotik; yapısal olarak makul çok az unsur var.
- 2: Ciddi yapısal sorunlar; sadece küçük unsurlar tesadüfen makul.
- 1: Neredeyse tamamen yanlış; asgari yapısal tutarlılık tesadüfen.
- 0: Tamamen yanlış; hiçbir yapısal tutarlılık yok.

Lütfen değerlendirmenizi döndürürken aşağıdaki JSON formatına kesinlikle uyun:

```
{  
  "Düğüm Doğruluğu": {"Score": (0–10), "Brief Reasoning": "..."},  
  "Edge_Accuracy": {"Score": (0–10), "Brief_Reasoning": "..."},  
  "Overall_Structural_Quality": {"Score": (0–10), "Brief_Reasoning": "..."}  
}
```

Şimdi, değerlendirmenize geçin:

Causal Reasoning Problem Context:
{problem_context}

Ground Truth Causal Graph Description (Evaluation Standard):
{reasoning with ground truth causal graph}

Model-generated Causal Graph Description (Evaluation Target):
{LLM's output for the problem (including causal diagram)}

Example of LLM-based Evaluation of Causal Graph Quality To clearly illustrate the LLM-based evaluation procedure used in our experiments, we provide a detailed example. Below, we demonstrate step-by-step how we objectively assessed the quality of causal graphs generated by different training methods (BaseModel (Llama-3.1-8B), and CDCR-SFT-Enhanced).

Detailed Evaluation Example

(1) Causal Reasoning Problem Context:

Imagine a self-contained, hypothetical world with only the following conditions, and without any unmentioned factors or causal relationships: Demand has a direct effect on supply and price. Yield per acre has a direct effect on supply. Supply has a direct effect on price. Demand is unobserved. The overall probability of increased supply is 60%. The probability of reduced supply and increased price is 25%. The probability of increased supply and increased price is 24%. Is the chance of increased price smaller when observing increased supply?

(2) Ground Truth Causal Graph (Evaluation Standard):

- Nodes:** Let V2 = yield per acre; V1 = demand; X = supply; Y = price.
- Edges:** V1 ->X, V2 ->X, V1 ->Y, X ->Y

(3) Model-generated Causal Graphs (Evaluation Targets):

BaseModel (Llama-3.1-8B) Raw Output:

To analyze the scenario, we need to construct a causal graph.

The causal graph for the scenario is as follows:

A (Demand) ->B (Supply)
B (Supply) ->C (Price)
B (Supply) ->D (Yield per acre)

From the given probabilities, we have the following conditional probability tables: [omit reasoning parts...]

CDCR-SFT-Enhanced (Llama-3.1-8B) Raw Output:

Alright, let me first review your input. Next, I will build a causal graph from the information provided, defining each node and clarifying how they interact. After that, I will detail the steps of causal inference, describing how I move from the causal graph to the final answer. To ensure clarity, I will begin by presenting the causal graph's structure, the meaning of each node, and their connections. Then, I will illustrate the inference process, leading up to the result.

Causal Graph:

First, here is the section on the causal graph nodes. For each node, I will list its ID, Name, and provide a brief description.

Nodes: Node ID: V1 Node Name: Demand Node Description: Unobserved variable representing demand. Directly influences both supply and price.

Node ID: Y Node Name: Price Node Description: Price level. Directly influenced by demand and supply.

Node ID: X Node Name: Supply Node Description: Supply level. Directly influenced by yield per acre and demand, and directly affects price.

Node ID: V2 Node Name: Yield per Acre Node Description: Yield per acre. Directly influences supply.

Next, I will explain how these nodes are linked in the causal graph, showing how information flows between them. I will go through each node in turn, indicating which nodes feed into it and which nodes it influences.

Edges: Node: V1 Inputs: N/A Outputs: ['X', 'Y']

Node: Y Inputs: ['V1', 'X'] Outputs: N/A

Node: X Inputs: ['V1', 'V2'] Outputs: ['Y']

Node: V2 Inputs: N/A Outputs: ['X']

[omit reasoning parts...]

Nedensel Akıl Yürütme Problemi Bağlamı:
{problem.bağlamı}

Zemin Gerçeği Nedensel Çizge Açıklaması (Değerlendirme Standardı):
{zemin gerçeği nedensel çizge ile akıl yürütme}

Model Tarafından Oluşturulan Nedensel Çizge Açıklaması (Değerlendirme Hedefi):
{BDM'nin problem için çıktı (nedensel diyagram dahil)}

BDM Tabanlı Nedensel Çizge Kalitesi Değerlendirme Örneği Deneylerimizde kullanılan BDM tabanlı değerlendirme prosedürü net bir şekilde açıklamak için ayrıntılı bir örnek sunuyoruz. Aşağıda, farklı eğitim yöntemleri (BaseModel (Llama-3.1-8B) ve CDCR-SFT-Enhanced) tarafından oluşturulan nedensel çizgelerin kalitesini objektif olarak nasıl değerlendirdiğimizi adım adım gösteriyoruz.

Ayrıntılı Değerlendirme Örneği

(1) Nedensel Akıl Yürütme Problemi Bağlamı:

Yalnızca aşağıdaki koşulların bulunduğu ve belirtilmeyen herhangi bir faktör veya nedensel ilişkinin olmadığı, kendi içinde kapalı, hipotetik bir dünya hayal edin: Talep, tedarik ve fiyat üzerinde doğrudan bir etkiye sahiptir. Dönüm başına verim, tedarik üzerinde doğrudan bir etkiye sahiptir. Tedarik, fiyat üzerinde doğrudan bir etkiye sahiptir. Talep gözlemlenmemiştir. Artan tedarikin genel olasılığı %60'tır. Azalan tedarik ve artan fiyat olasılığı %25'tir. Artan tedarik ve artan fiyat olasılığı %24'tür. Artan tedarik gözlemlendiğinde, artan fiyat şansı daha mı düşüktür?

(2) Zemin Gerçeği Nedensel Çizgesi (Değerlendirme Standardı):

- Düğümler:** V2 = dönüm başına verim; V1 = talep; X = tedarik; Y = fiyat olsun.
- Kenarlar:** V1 ->X, V2 ->X, V1 ->Y, X ->Y

(3) Model Tarafından Oluşturulan Nedensel Çizgeler (Değerlendirme Hedefleri): Temel Model (Llama-3.1-8B) Ham Çıktısı:

Senaryoyu analiz etmek için nedensel bir çizge oluşturmak gerekiyor.

Senaryo için nedensel çizge aşağıdaki gibidir:

A (Talep) ->B (Tedarik)
B (Tedarik) ->C (Fiyat)
B (Tedarik) ->D (Dönüm başına verim)

Verilen olasılıklardan aşağıdaki koşullu olasılık tablolarına sahibiz: [akıl yürütme kısımları çıkarıldı...]

CDCR-SFT-Geliştirilmiş (Llama-3.1-8B) Ham Çıktı:

Pekala, öncelikle girdinizi gözden geçirmeme izin verin. Ardından, sağlanan bilgilerden bir nedensel çizge oluşturacak, her düğümü tanımlayacak ve nasıl etkileşimde bulunduklarını açıklayacağım. Daha sonra, nedensel çıkarım adımlarını detaylandıracak, nedensel çizgiden nihai cevaba nasıl ulaştığımı açıklayacağım. Netliği sağlamak adına, nedensel çizgenin yapısını, her düğümün anlamını ve bağlantılarını sunarak başlayacağım. Ardından, çıkarım sürecini ve sonuca giden adımları göstereceğim.

Nedensel Çizge:

İlk olarak, nedensel çizge düğümleri bölümü aşağıdadır. Her düğüm için, Düğüm Kimliğini, Düğüm Adını ve kısa bir Düğüm Açıklamasını listeleyeceğim.

Düğümler: Düğüm Kimliği: V1 Düğüm Adı: Talep Düğüm Açıklaması: Gözlemlenmeyen talep değişkeni. Hem tedariki hem de fiyatını doğrudan etkiler.

Düğüm Kimliği: Y Düğüm Adı: Fiyat Düğüm Açıklaması: Fiyat seviyesi. Talep ve tedarikten doğrudan etkilenir.

Düğüm Kimliği: X Düğüm Adı: Tedarik Düğüm Açıklaması: Tedarik seviyesi. Dönüm başına verim ve talepten doğrudan etkilenir, ayrıca fiyatını doğrudan etkiler.

Düğüm Kimliği: V2 Düğüm Adı: Dönüm başına verim Düğüm Açıklaması: Dönüm başına verim. Tedariki doğrudan etkiler.

Ardından, bu düğümlerin nedensel çizgede nasıl bağlandığını ve aralarındaki bilgi akışını açıklayacağım. Her bir düğüm sırasıyla ele alacak, hangi düğümlerin onu etkilediğini ve hangi düğümleri etkilediğini belirteceğim.

Kenarlar: Düğüm: V1 Girdiler: N/A Çıktılar: ['X', 'Y']

Düğüm: Y Girdiler: ['V1', 'X'] Çıktılar: N/A

Düğüm: X Girdiler: ['V1', 'V2'] Çıktılar: ['Y']

Düğüm: V2 Girdiler: Yok Çıktılar: ['X']

[akıl yürütme kısımları çıkarıldı...]

(4) LLM-based Evaluation Results:

Below are the LLM-generated evaluation results, strictly adhering to our defined scoring criteria and JSON format described in Appendix B.4.

BaseModel Evaluation:

```
1 {
2   "Node_Accuracy": {
3     "Score": 6,
4     "Brief_Reasoning": "The model-generated graph correctly identifies the core nodes (Demand, Supply, Price) but misidentifies the relationship with Yield per acre, which should be a direct effect on Supply rather than an edge from Supply to Yield."
5   },
6   "Edge_Accuracy": {
7     "Score": 5,
8     "Brief_Reasoning": "The model-generated graph has the correct edges for Demand to Supply and Supply to Price, but incorrectly includes an edge from Supply to Yield instead of Yield to Supply, leading to significant errors in edge identification."
9   },
10  "Overall_Structural_Quality": {
11    "Score": 5,
12    "Brief_Reasoning": "The structure is somewhat coherent with the core relationships but contains critical misrepresentations, particularly regarding the directionality of edges and the inclusion of Yield per acre, which affects the overall logical flow."
13  }
14 }
```

CDCR-SFT-Enhanced Evaluation:

```
1 {
2   "Node_Accuracy": {
3     "Score": 10,
4     "Brief_Reasoning": "All nodes (Demand, Price, Supply, Yield per Acre) are perfectly identified and accurately described."
5   },
6   "Edge_Accuracy": {
7     "Score": 10,
8     "Brief_Reasoning": "All edges are correctly identified with the correct directions, reflecting the causal relationships as per the Ground Truth."
9   },
10  "Overall_Structural_Quality": {
11    "Score": 10,
12    "Brief_Reasoning": "The structure of the model-generated causal graph perfectly matches the Ground Truth, with clear and coherent relationships among the nodes."
13  }
14 }
```

The above example explicitly illustrates the exact procedure and transparency of our evaluation methodology, starting from raw model outputs, extracting structured causal graphs, and finally obtaining standardized LLM-based scores.

B.5 Case Study Supplementary

To demonstrate the practical effectiveness of CDCR-SFT, we present a case illustration on the complete execution flow for a counterfactual reasoning query from WIQA that requires understanding complex causal mechanisms in Fig. 6. The selected query exemplifies the challenges that CDCR-SFT addresses: *"Process steps: An adult frog spawns eggs in water | These eggs hatch into tadpoles and continue to live in the water | The tadpoles grow developing external gills and a longer tail | The tadpole begins storing food in the tail | The tadpole develops hind legs and lives off food stored in its tail | The front legs appear and the tadpoles tail shortens | The tadpole now looks like a young frog | The young frog continues to grow to adulthood. Question: Will the absence of adult frogs laying eggs cause more frog population, less frog population, or have no effect?" Choice A: more; Choice B: less; Choice C: no effect*", which demands exact node identification, edge direction, and graph-based inference to trace effects without adding false links. The causal DAG construction phase breaks down the scenario into nodes and edges. The reasoning phase builds on the DAG to trace paths step by step following the reasoning goal. Finally, we got the answer.

(4) BDM tabanlı Değerlendirme Sonuçları:

Aşağıda, Ek B.4'te açıklanan tanımlanmış puanlama kriterlerimize ve JSON formatına kesinlikle uygun olarak BDM tarafından oluşturulan değerlendirme sonuçları bulunmaktadır.

Temel Model Değerlendirmesi:

```
1 {
2   "Düğüm_Doğruluğu": {
3     "Puan": 6,
4     "Kısa_Gerekçe": "Model tarafından oluşturulan grafik, temel düğümleri (Talep, Tedarik, Fiyat) doğru bir şekilde tanımlamaktadır ancak dönüm başına verim ile olan ilişkisi yanlış belirlemiştir; bu, Tedarikten Verime bir kenar yerine Tedarik üzerinde doğrudan bir etki olmalıdır."
5   },
6   "Kenar_Doğruluğu": {
7     "Puan": 5,
8     "Kısa_Gerekçe": "Model tarafından oluşturulan grafik, Talep'ten Tedarik'e ve Tedarik'ten Fiyat'a doğru kenarlara sahiptir ancak Dönüm Başına Verim'den Tedarik'e yerine Tedarik'ten Dönüm Başına Verim'e yanlış bir kenar içerecekken kenar tanımlamasında önemli hatalara yol açmaktadır."
9   },
10  "Genel_Yapısal_Kalite": {
11    "Puan": 5,
12    "Kısa_Gerekçe": "Yapı, temel ilişkilerle bir miktar tutarlıdır ancak özellikle kenarların yönü ve dönüm başına verimin dahil edilmesi konusunda kritik yanlış temsiller içerir; bu da genel mantıksal akışı etkiler."
13  }
14 }
```

CDCR-SFT-Gelişmiş Değerlendirme:

```
1 {
2   "Düğüm_Doğruluğu": {
3     "Puan": 10,
4     "Kısa_Gerekçe": "Tüm düğümler (Talep, Fiyat, Tedarik, Dönüm Başına Verim) mükemmel bir şekilde tanımlanmış ve doğru bir şekilde açıklanmıştır."
5   },
6   "Kenar_Doğruluğu": {
7     "Puan": 10,
8     "Kısa_Gerekçe": "Tüm kenarlar, Zemin Gerçeği'ne göre nedensel ilişkileri yansitan doğru yönlerle doğru bir şekilde tanımlanmıştır."
9   },
10  "Genel_Yapısal_Kalite": {
11    "Puan": 10,
12    "Kısa_Gerekçe": "Model tarafından oluşturulan nedensel çizgenin yapısı, düğümler arasındaki açık ve tutarlı ilişkilerle Zemin Gerçeği'ne mükemmel bir şekilde uymaktadır."
13  }
14 }
```

Yukarıdaki örnek, ham model çıktılarından başlayarak, yapılandırılmış nedensel çizgelerin çıkarılmasına ve son olarak standartlaştırılmış BDM tabanlı puanların elde edilmesine kadar değerlendirme metodolojimizin prosedürüne ve şeffaflığını açıkça göstermektedir.

B.5 Vaka Çalışması Tamamlayıcısı

CDCR-SFT'nin pratik etkinliğini göstermek amacıyla, Şekil 6'da karmaşık nedensel mekanizmaları anlamayı gerektiren WIQA'dan alınan karşıolgusal bir akıl yürütme sorgusu için eksiksiz yürütme akışına ilişkin bir vaka örneği sunuyoruz. Seçilen soru, CDCR-SFT'nin ele aldığı zorlukları örneklemektedir: *"Süreç adımları: Yetişkin bir kurbağa suya yumurta bırakır | Bu yumurtalar kurbağa yavrularına dönüşür ve suda yaşamaya devam eder | Kurbağa yavruları dış solungaçlar geliştirerek ve daha uzun bir kuyrukla büyür | Kurbağa yavrusu kuyruğunda yiyecek depolamaya başlar | Kurbağa yavrusu arka bacaklar geliştirir ve kuyruğunda depoladığı yiyeceleyle yaşı | Ön bacaklar belirir ve kurbağa yavrusunun kuyruğu kısalır | Kurbağa yavrusu artık genç bir kurbağaya benzer | Genç kurbağa yetişkinliğe kadar büyümeye devam eder. Soru: Yetişkin kurbağaların yumurta bırakmaması, kurbağa popülasyonunda artış mı, azalıtı mı neden olur, yoksa hiçbir etkisi olmaz mı?" Seçenek A: daha fazla; Seçenek B: daha az; Seçenek C: etkisi yok*", bu durum, yanlış bağlantılar eklenmeden etkileri izlemek için kesin düzgün tanımlama, kenar yönü ve çizge tabanlı çıkarım gerektirir. Nedensel YADÇ oluşturma aşaması, senaryoyu düğümlere ve kenarlara ayırır. Akıl yürütme aşaması, akıl yürütme hedefini takip ederek adımları izlemek için YADÇ üzerine inşa edilir. Sonunda, cevabı aldık.

Alright, let me first review your input. Next, I will build a causal graph from the information provided, defining each node and clarifying how they interact. After that, I will detail the steps of causal inference, describing how I move from the causal graph to the final answer. To ensure clarity, I will begin by presenting the causal graph's structure, the meaning of each node, and their connections. Then, I will illustrate the inference process, leading up to the result.

Causal Graph:

First, here is the section on the causal graph nodes. For each node, I will list its ID, Name, and provide a brief description.

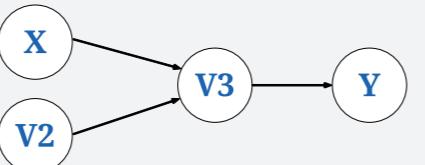
Nodes:

Node ID: X\nNode Name: Adult Frog Reproduction\nNode Description: The ability of adult frogs to lay eggs, which directly influences the number of tadpoles and subsequently the frog population.
 Node ID: V2\nNode Name: Tadpole Survival\nNode Description: The survival rate of tadpoles, affected by environmental factors and parental traits.
 Node ID: V3\nNode Name: Tadpole Development\nNode Description: The developmental stage of tadpoles, influenced by their environment and parental traits through egg transmission.
 Node ID: Y\nNode Name: Frog Population\nNode Description: The total number of frogs in the population, determined by survival and reproduction rates.

Next, I will explain how these nodes are linked in the causal graph, showing how information flows between them. I will go through each node in turn, indicating which nodes feed into it and which nodes it influences.

Edges:

```
Node: X\nInputs: N/A\nOutputs: ['V3']
\nNode: V2\nInputs: ['X']\nOutputs: ['V3']
Node: V3\nInputs: ['X', 'V2']\nOutputs: ['Y']
Node: Y\nInputs: ['V3']\nOutputs: N/A
```



Based on the current provided input information, the inference goal is that "Determine whether the absence of adult frogs laying eggs causes an increase, decrease, or no effect in the frog population."

Given this goal, the following describes the **causal inference process**.

Causal Inference Process: The causal graph shows that adult frog reproduction (X) directly affects tadpole development (V3) and indirectly influences tadpole survival (V2). When X=0 (no egg-laying), V3=0 (no tadpoles), and V2=0 (no survivors). Since Y is fully determined by V3 and V2 ($Y = V3 \wedge V2$), $Y=0$. The absence of eggs eliminates all potential frogs, as tadpoles require eggs to develop. This causal pathway confirms that preventing adult frogs from laying eggs (X=0) reduces the population to zero, making the answer 'less.'

As a result of this causal inference process, I will reply with the **answer choice B.**"

Figure 6: CDCR-SFT inference trace demonstrating the output integrating causal DAG, reasoning path, and answer for a counterfactual reasoning query.

Pekala, öncelikle girdinizi gözden geçirmeme izin verin. Ardından, sağlanan bilgilerden bir nedensel çizge oluşturacağım, her düğümü tanımlayacak ve nasıl etkileşime girdiklerini açıklayacağım. Bundan sonra, nedensel çıkarım adımlarını detaylandıracıım; nedensel çizgeden nihai cevabı nasıl geçtiğini anlatacağım. Netliği sağlamak için, öncelikle nedensel çizgenin yapısını, her düğümün anlamını ve bağlantılarını sunarak başlayacağım. Daha sonra, sonuçlara yol açan çıkışın sürecini göstereceğim.

Nedensel Çizge:

İlk olarak, nedensel çizge düğümlerine ilişkin bölüm aşağıdadır. Her düğüm için Kimliğini, Adını ve kısa bir açıklamasını listeleyeceğim.

Düğümler:

Düğüm Kimliği: X \ nDüğüm Adı: Yetişkin Kurbağa Üremesi \ nDüğüm Açıklaması: Yetişkin kurbağaların yumurta bırakma yeteneği, bu durum iribaşların sayısını ve dolayısıyla kurbağa popülasyonunu doğrudan etkiler.

Düğüm Kimliği: V2 \ nDüğüm Adı: İribaş Yaşam Oranı \ nDüğüm Açıklaması: Çevresel faktörlerden ve ebeveyn özelliklerinden etkilenen iribaşların yaşam oranıdır.

Düğüm Kimliği: V3 \ nDüğüm Adı: İribaş Gelişimi \ nDüğüm Açıklaması: Çevrelerinden ve yumurta transferi yoluyla ebeveyn özelliklerinden etkilenen iribaşların gelişim aşamasıdır.

Düğüm Kimliği: Y \ nDüğüm Adı: Kurbağa Popülasyonu \ nDüğüm Açıklaması: Yaşam ve üreme oranları tarafından belirlenen popülasyondaki toplam kurbağa sayısıdır.

Daha sonra, bu düğümlerin nedensel çizgede nasıl bağlı olduğunu, aralarındaki bilgi akışını göstererek açıklayacağım. Her bir düğümü sırayla inceleyeceğim, hangi düğümlerin ona girdi sağladığını ve hangi düğümleri etkilediğini belirteceğim.

Kenarlar:

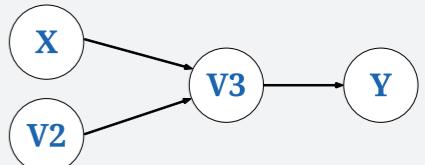
```
Düğüm: X \ nGirdiler: Yok \ nÇıktılar: ['V3']
\ nDüğüm: V2 \ nGirdiler: ['X'] \ nÇıktılar: ['V3']
Düğüm: V3 \ nGirdiler: ['X', 'V2'] \ nÇıktılar: ['Y']
Düğüm: Y \ nGirdiler: ['V3'] \ nÇıktılar: Yok
```

Mevcut girdi bilgilerine dayanarak, çıkışım hedefi "Yetişkin kurbağaların yumurta bırakmamasının kurbağa popülasyonunda artışa, azalışa veya hiçbir etkiye neden olup olmadığını belirlemek." şeklindeki.

Bu hedef doğrultusunda, nedensel çıkarım süreci aşağıda açıklanmaktadır.

Nedensel Çıkarım Süreci: Nedensel çizge, yetişkin kurbağa üremesinin (X) doğrudan iribaş gelişimini (V3) ve dolaylı olarak iribaş hayatı kalmasını (V2) etkilediğini göstermektedir. $X = 0$ (yumurtlama yok) olduğunda, $V3 = 0$ (iribaş yok) ve $V2 = 0$ (hayatta kalan yok) olur. Y , $V3$ ve $V2$ tarafından tamamen belirlendiği için ($Y = V3 \wedge V2$), $Y = 0$ 'dır. Yumurtaların yokluğu, iribaşların gelişmek için yumurtalarla ihtiyaç duyması nedeniyle tüm potansiyel kurbağaları ortadan kaldırır. Bu nedensel yol, yetişkin kurbağaların yumurta bırakmasını engellemenin ($X = 0$) popülasyonu sıfıra indirdiğini ve cevabı 'daha az' yaptığını doğrular.

Bu nedensel çıkarım süreci sonucunda, B seçeneğiyle yanıt vereceğim."



Şekil 6: Nedensel Yönlendirilmiş Dönüşel Olmayan Çizelge, akıl yürütme yolu ve karşı olgusal bir akıl yürütme sorusu için cevabı entegre eden CDR-SFT çıkarım izini gösteriyor.