CGNTM: Unsupervised Causal Topic Modeling with LLMs and Nonlinear Causal GNNs

Peixuan Men[a], Longchao Wang[b], Aihua Li[c] and Xiaoli Tang[d]
Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Keywords: Causal Topic Modeling, Unsupervised Learning, Neural Causal Modeling, Graph Neural Networks.

Abstract: We propose CGNTM, a fully unsupervised causal topic model that integrates large language models (LLMs) with neural causal inference. Unlike conventional and supervised topic models, CGNTM learns both hierarchical topics and their directed causal relations directly from raw text, without requiring labeled data. The framework leverages LLM-based prompt extraction to identify salient keywords and candidate causal pairs, which are refined through differentiable Directed Acyclic Graph (DAG) learning and modeled via a nonlinear structural causal model (SCM). A directionally masked graph neural network (GNN) propagates information strictly along causal edges, while a Wasserstein Generative Adversarial Network (GAN) enforces semantic consistency under counterfactual interventions via BERT-based regularization. This combination enables the model to not only discover coherent and diverse topics but also uncover interpretable causal relationships among them. The architecture supports hierarchical topic organization by clustering fine-grained terms into broader themes and modeling cross-level dependencies through dual-layer message passing. Experimental results demonstrate that CGNTM outperforms state-of-the-art models in topic quality and causal interpretability. Ablation studies confirm the essential role of each component-LLM-guided extraction, nonlinear SCM, directional GNN propagation, and adversarial training-in contributing to both causal accuracy and topic coherence. The proposed framework opens new directions for unsupervised causal discovery in text, offering transformative potential in domains where understanding why certain topics co-occur is as crucial as identifying what they are.

1 INTRODUCTION

Topic modeling is a vital tool in natural language processing for uncovering hidden themes in large text corpora. Classical models like Latent Dirichlet Allocation (LDA) summarize documents into interpretable topics, supporting tasks such as classification and retrieval, but rely on bag-of-words, assume independence, and ignore semantic dependencies, limiting interpretability and omitting concept relationships (Morstatter and Liu, 2018). Recent Neural Topic Models (NTMs) leverage deep generative networks for flexible inference, enhancing coherence through contextualized embeddings or external knowledge (Shen et al., 2021). However, they capture only statistical co-occurrence, not causal relationships among topics, hindering interpretability and the ability to answer "why" questions from text data.

Recent efforts integrate causality, such as the supervised Causal Relationship-Aware Neural Topic Model (CRNTM) (Tang et al., 2024), which uses Structural Causal Models (SCMs) to uncover topic-label links in a Directed Acyclic Graph (DAG). This improves structure and quality but requires supervision. Discovering causal relations in unlabeled corpora, particularly with hierarchical organization, remains an open challenge (Lagemann et al., 2023).

This paper addresses unsupervised causal topic discovery: identifying hierarchical topics and inferring a DAG of their causal relationships from raw text without supervision. This tackles the intertwined challenges of multi-granularity topic extraction and causal graph inference using statistical patterns and semantic knowledge for interpretable structures.

We propose the Causal Graph Neural Topic Model (CGNTM), integrating LLMs, causal graph learning,

[a] https://orcid.org/0009-0002-2630-3838
[b] https://orcid.org/0009-0009-1387-3517
[c] https://orcid.org/0000-0001-6742-3268
[d] https://orcid.org/0000-0001-6946-3482

Men, P., Wang, L., Li, A. and Tang, X.
CGNTM: Unsupervised Causal Topic Modeling with LLMs and Nonlinear Causal GNNs.
DOI: 10.5220/0013708200004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 275-285

and a GNN in an unsupervised pipeline. It extracts keywords and causal pairs via LLMs, refines them through NOTEARS (Zheng et al., 2018) to enforce DAG constraints, models nonlinear causality with a neural SCM and directional GNN, and refines results via a WGAN with BERT-based consistency for counterfactual generation, enhancing topic coherence and interpretability.

In summary, this paper introduces CGNTM as a novel unsupervised framework for causal topic discovery. Its key contributions are as follows:

- The first unsupervised topic modeling approach integrating LLMs to extract causal relationships and construct a causal topic graph from unlabeled corpora.
- A novel architecture combining a nonlinear neural SCM with a masked directional GNN for cause-to-effect propagation.
- Hierarchical causal structure via two-layer GNN and clustering for multi-granular interpretation.
- Adversarial counterfactual training with WGAN and BERT consistency to improve coherence and prevent spurious links.

2 RELATED WORK

2.1 Neural and Hierarchical Topic Modeling Approaches

Topic modeling has evolved from classical probabilistic models to deep learning-based methods. Traditional approaches like Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its hierarchical extensions (Miao et al., 2016) infer topics as latent variables under assumptions such as word exchangeability and topic independence, which limit expressiveness and coherence.

NTMs address these by incorporating variational autoencoders or transformers for flexible representations. Key examples include ProdLDA (Srivastava and Sutton, 2017), Embedded Topic Model (ETM) (Dieng et al., 2020) , and contextualized models (Venugopalan and Gupta, 2022), which use learned distributions and semantic embeddings to boost interpretability. Recent advances integrate graphs or knowledge, such as BERTopic (Grootendorst, 2022) for embedding-based topics and BERT-Flow-VAE (Liu et al., 2022) for calibrated embeddings in VAE modeling.

However, these methods still treat topic relations as statistical correlations or taxonomic hierarchies rather than causal dependencies.

2.2 Causal Discovery and Structural Causal Models in NLP

Causal discovery is gaining traction in NLP, especially topic modeling. SCMs formalize relations via functional equations like , where are parents in a DAG (Yang et al., 2022; Pawlowski et al., 2020), though unsupervised text applications are limited.

Works like CRNTM incorporate SCMs to link latent topics and labels , but require supervision. CausalVAE adds a causal layer in VAEs for latent DAGs without strong supervision (Panwar et al., 2020; Yang et al., 2021), yet focuses on disentangled factors, not textual topics (Wu et al., 2024). Few explore fully unsupervised causal discovery in NLP (Prostmaier et al., 2025), often relying on labels or disentanglement rather than interpretable themes.

2.3 GNNs for Causal Inference and Topic Modeling

GNNs model relational patterns effectively. In causal discovery, DAG-GNN combines GNNs with VAEs for DAG learning from data, optimizing acyclicity constraints (Yu et al., 2019; Park and Kim, 2023). For topic modeling, GNNs propagate information via graphs, as in GNTM (Shen et al., 2021) and GraphBTM (Zhu et al., 2018) using co-occurrence statistics. However, integrating GNN-driven causal learning with topics is underexplored (Gao et al., 2024; Behnam and Wang, 2024) methods treat GNNs as auxiliaries, not embedding them in causal pipelines. No work jointly learns topics and causal structures end-to-end unsupervised.

2.4 Large Language Models for Zero-Shot Keyword and Relation Extraction

LLMs demonstrate strong zero-shot capabilities in extracting keywords and semantic relations from text (Rana et al., 2024; Chen et al., 2023). Using natural-language prompts, LLMs can identify salient terms and hypothesize causal connections without task-specific training, making them valuable for uncovering candidate causal pairs in unlabeled corpora. Recent works leverage LLMs for weakly supervised causal discovery by querying plausible relations or generating pseudo-labels. However, standalone LLM extraction lacks consistency across documents and fails to yield a structured topic model.

These limitations underscore the need for CGNTM, which integrates LLM-driven knowledge

extraction with neural causal inference. Unlike conventional models such as BERTopic, which lack causal interpretation, or CRNTM, which requires supervision, CGNTM enables fully unsupervised causal topic discovery. It combines zero-shot prompt-based extraction with a nonlinear GNN-based causal inference module, facilitating interpretable topic modeling and causal structure learning without labeled data. This positions CGNTM as a novel contribution at the intersection of NLP, causality, and deep learning.

3 METHODS

We propose a multi-stage framework (Figure 1) that combines LLM-based semantic parsing with neural causal graph modeling to uncover latent topics and their causal relations. The pipeline proceeds as follows: (1) extract candidate keywords and causal pairs from text using a LLM; (2) construct a global directed causal graph over these candidates and refine it via NOTEARS continuous optimization to enforce acyclicity; (3) instantiate a SCM on this graph with neural functional mechanisms; (4) perform directed message passing with a GNN that masks reverse (anti-causal) edges; (5) extend the GNN to a dual-layer architecture for hierarchical topic representation; (6) generate counterfactual textual data by intervening on the learned causal topic variables and using a conditional WGAN with a BERT-based consistency module; and (7) train all components under a joint objective. The subsequent sections elaborate on each stage in detail.

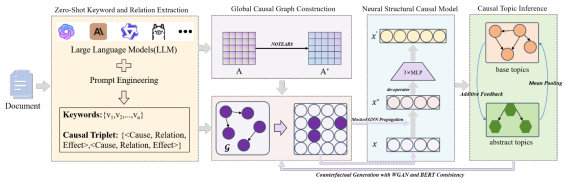

Figure 1: The overall multi-stage framework of CGNTM.

As illustrated in Figure 1, our framework seamlessly integrates symbolic knowledge extraction with neural causal modeling. The following subsections detail each component's technical implementation and theoretical foundations.

3.1 Keyword and Causal Pair Extraction via LLMs

The first stage extracts candidate topic variables (keywords) and causal links from an unlabeled corpus using a large language model (LLM). For each document, the LLM is prompted to output: (a) salient keywords summarizing its content, and (b) any explicit cause-effect pairs (e.g., "X causes Y") mentioned in the text. Prior studies show LLMs can identify high-quality keyphrases and infer causal relations in a zero-shot manner. Building on this, we aggregate each document's extracted keywords and directed pairs into a global set of topic terms and candidate edges. Each edge may carry a confidence score based on frequency or LLM-reported likelihood. This unsupervised parsing injects semantic knowledge into the pipeline, providing a plausible initial causal graph structure.

3.2 Global Causal Graph Construction and NOTEARS Optimization

We next consolidate the LLM-derived keywords and causal pairs into a unified global causal topic graph. This graph is represented by an adjacency matrix, where each entry indicates a directed edge between topics. The initial structure is constructed from the candidate edges identified in Section 3.1, with weights set either uniformly or based on LLM-provided confidence scores. Since these raw connections may contain cycles or noisy links, we refine the graph by solving a structure learning problem with an acyclicity constraint.

To this end, we use the NOTEARS method, which casts DAG learning as a smooth optimization problem. It learns a weighted adjacency matrix that minimizes a given loss while ensuring the graph remains acyclic through a differentiable constraint based on the matrix exponential formulation. Specifically, we impose:

$$h(A) = \mathrm{Tr}\left[\exp(A \circ A)\right] - m = 0 \qquad (1)$$

where $A \circ A$ denotes the elementwise square and $m = |V|$ is the number of nodes. Intuitively, $\mathrm{Tr}\left(\exp(A \circ A)\right) = m$ if and only if A has no cyclic dependencies. This hard equality constraint can be handled via a penalty or augmented Lagrangian method during optimization. We thus solve:

$$\min_{A,\Theta} \mathcal{L}_{\text{score}}(A,\Theta) \quad \text{s.t.} \quad h(A) = 0 \qquad (2)$$

where Θ denotes other model parameters (discussed later) and $\mathcal{L}_{\text{score}}$ is a structure score or loss. We initialize A with the LLM-extracted graph and iteratively update it to reduce $\mathcal{L}_{\text{score}}$ while applying the DAG constraint (using gradient-based optimization). This yields a global causal topic graph that best explains the data without cycles. By casting structure

discovery as continuous optimization, we avoid brute-force search over graphs and can efficiently handle the moderate number of topic nodes. The resulting adjacency A (with learned weights) will serve as the backbone for our neural causal model.

3.3 Neural Structural Causal Model (NSCM)

Given the DAG learned in section 3.2, we define a structural causal model to describe how each topic variable is generated as a function of its causes. Following the Pearlian framework, an SCM consists of a set of structural equations $X_i = f_i(\mathrm{PA}_i, N_i)$ for each node X_i (representing topic v_i), where PA_i denotes the set of parent variables (direct causes of X_i in G) and N_i is an exogenous noise term (independent for each i) capturing unmodeled variation.

In our Neural SCM, we parameterize each structural function as a neural network, allowing for complex non-linear causal relationships between topics. In particular, for each directed edge $v_j \to v_i$ in the graph, the causal influence of topic v_j on topic v_i is modeled by learnable weights within the neural function f_i. Formally, let z_i be a latent representation or "activation" of topic v_i. We define:

$$z_i = f_i\left(\{z_j : v_j \in \mathrm{PA}_i\}\right) + n_i \tag{3}$$

where $n_i \sim \mathcal{N}(0, \sigma^2)$ (or another suitable noise distribution) and f_i is implemented as a multilayer perceptron (MLP). This yields a deep non-linear structural equation model(SEM) that generalizes traditional linear causal models.

The collection $\{f_i\}_{i=1}^m$ together with adjacency A defines a joint distribution $P(X_1, \ldots, X_m)$ over topic variables – essentially a causal topic model. Importantly, because A is acyclic, we can sample from this model by ancestral sampling (order the topics topologically and sample each X_i from $f_i(\mathrm{PA}_i)$). The Neural SCM grants our model the capacity to capture complex interactions (e.g., a cause may affect an effect in a highly non-linear or context-dependent way) (Zečević et al., 2021). It also enables interventions: we can apply the do-operator $do(X_i = x^*)$ by clamping z_i to a chosen value and forward-simulating the rest of the model, to infer causal effects or counterfactual topic configurations.

In summary, at this stage we have a parameterized causal model $M = (A, \{f_i\})$ that explains how topics cause one another. The parameters of f_i (and any distributional parameters of N_i) will be learned from data, typically by maximizing the likelihood of observed topic occurrences in documents.

3.4 Directionally Masked GNN Propagation

To leverage the learned causal graph during both training and inference, we implement a directionally masked Graph Attention Network (GAT) that propagates information strictly along directed edges (Wang et al., 2025; Liu et al., 2022). Unlike conventional GATs that allow bidirectional attention, our model enforces causal directionality to align with the causal semantics of the structural causal model (SCM).

We implement a directionally masked GNN on the directed acyclic graph (DAG) learned in Section 3.2 (Kaushik et al., 2019). The DAG is represented by an adjacency matrix A, where A_{ji} denotes the weight of the directed edge from node j to node i, with $A_{ji} = 0$ for nonexistent or directionally excluded edges (i.e., edges not present in the causal graph). Each topic node v_i is associated with a hidden state vector $h_i^{(l)}$ at layer l of the GNN (with $h_i^{(0)}$ initialized from node features, e.g., textual embedding of the keyword or an initial topic score in a document). For each topic node v_i, we compute attention weights only over its parent nodes, defined as $\mathrm{PA}_i = \{j | A_{ji} > 0\}$:

$$\alpha_{ji} = \mathrm{softmax}(\mathrm{LeakyReLU}(a^T [W^{(l)} \cdot h_j^{(l)} \parallel W^{(l)} \cdot h_i^{(l)}])) \tag{4}$$

Then the layer-wise update for node i is defined as:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathrm{PA}_i} \alpha_{ji} \cdot W^{(l)} \cdot h_j^{(l)}\right) \tag{5}$$

where $\parallel$ denotes concatenation, a is a learnable attention vector, and $W^{(l)}$ are layer-specific weight matrices, and $\sigma(\cdot)$ is a non-linear activation function. The attention coefficients α_{ji} serve as dynamic, learned weights on edge $j \to i$, capturing causal influence adaptively based on node features.

In matrix form, the propagation can be expressed as:

$$H^{(l+1)} = \sigma\left(A \cdot (H^{(l)} \cdot W^{(l)})\right) \tag{6}$$

Where $H^{(l)}$ is the matrix of all node hidden states at layer l, and A is the masked adjacency matrix from Section 3.2, ensuring that information propagates only along the causal directions defined by the DAG. By stacking L such layers (where L is the number of GNN layers, typically set to 2–4 based on the graph depth and empirical tuning), each node's representation $h_i^{(L)}$ captures information from its causal ancestors up to L hops away. This directed GNN aligns

naturally with the semantics of the SCM: messages flow from causes to effects, mirroring how interventions or changes propagate in the causal graph. We use this GNN both to encode observational data into latent topic representations and to simulate the spread of causal influence in the topic graph.

Unlike standard graph convolutions that mix incoming and outgoing messages, our GNN restricts propagation to incoming edges defined by PA_i, implementing causal message passing. While inspired by generative graph models such as DAG-GNN, our architecture enforces strict temporal directionality, ensuring information flows from cause to effect. The directional GNN operates jointly with the neural SCM (Section 3.3) to enable efficient inference, where the NSCM defines the generative process z_i, specifying how topic variables are generated based on their causal parents, while the GNN implements the inference process by computing topic representations from observed data while respecting causal constraints. During training, the GNN parameters $\{W^{(l)}, a\}$ are learned jointly with the SCM functions $\{f_i\}$ as part of the unified optimization framework (Section 3.7), ensuring the inference network is consistent with the generative causal model. Given partial topic activations, the GNN infers missing variables via causal propagation. Under interventions, it rapidly computes perturbed topic representations, supporting counterfactual reasoning critical for applications.

3.5 Hierarchical Topic Modeling with Dual-Layer GNN

Real-world topics often exhibit hierarchical structure, with fine-grained concepts nested within broader thematic categories. To capture this hierarchy, we extend our single-layer causal topic model to a dual-layer architecture that explicitly represents both micro-level topics and macro-level themes. Our hierarchical approach consists of three sequential steps: bottom-up aggregation, horizontal propagation, and top-down refinement.

We introduce an upper layer of abstract topics that group together semantically related base topics. Concretely, suppose we partition V into K clusters $\{C_1, \ldots, C_K\}$ (each C_k is a subset of the base topics that constitute a higher-level theme). These clusters could be obtained by heuristic clustering of the keyword embeddings or even by another pass of LLM-based grouping. We then create K new nodes $\{\tilde{v}_1, \ldots, \tilde{v}_K\}$ representing the abstract topics (one per cluster). We connect each base topic node v_i to its abstract parent $\tilde{v}_k$ (if $v_i \in C_k$) via an undirected link, and we also allow directed causal edges among the abstract nodes themselves (induced by the base-level DAG: e.g., if some $v_i \in C_a$ causes $v_j \in C_b$, we add a directed edge $\tilde{v}_a \to \tilde{v}_b$ between abstract topics).

This yields a two-layer graph: Layer 1 comprises base topics V with directed edges E (from our learned A). Layer 2 comprises abstract topics $\tilde{V}$ with directed edges among them. Additionally, bipartite connections link each $\tilde{v}_k$ to all $v_i \in C_k$.

We then design a two-tier message passing scheme:

3.5.1 Bottom-up Aggregation

In the bottom-up aggregation step, each abstract topic node k computes an initial state as the mean of its member topic embeddings:

$$\widetilde{h}_k^{(0)} = \frac{1}{|C_k|} \sum_{v_i \in C_k} h_i^{(L)} \tag{7}$$

using the final embeddings $h_i^{(L)}$ from the base GNN layer as input. This step produces a representation for the higher-level topic as a composition of its subtopics.

3.5.2 Horizontal Propagation at the Abstract Level

Next, in the horizontal propagation step, we apply another GNN on the abstract topic graph (layer 2) for L' steps, using the directed edges among $\widetilde{V}$. This is analogous to section 3.4 but on the smaller graph of K nodes.For each abstract topic node $\widetilde{v}_a$, we compute attention weights only over its parent nodes:

$$\alpha_{mk} = \mathrm{softmax}(\mathrm{LeakyReLU}(a^T \\ [W^{(l)} \cdot \widetilde{h}_m^{(l)} \parallel W^{(l)} \cdot \widetilde{h}_k^{(l)}])) \tag{8}$$

The layer-wise update for node $\widetilde{v}_K$ is:

$$\widetilde{h}_k^{(l+1)} = \sigma\left(\sum_{\widetilde{v}_m \in \mathrm{PA}(\widetilde{v}_k)} \alpha_{mk} \cdot W^{(l)} \cdot \widetilde{h}_m^{(l)} \right), \\ l = 0, \ldots, L' - 1 \tag{9}$$

where $\mathrm{PA}(\widetilde{v}_k)$ are abstract parent nodes of $\widetilde{v}_k$ in the abstract DAG. α_{mk} is the attention coefficient between parent $\widetilde{v}_m$ and target $\widetilde{v}_k$. This yields refined high-level topic embeddings $\widetilde{h}_k^{(L')}$ that capture how broad themes influence each other.

3.5.3 Top-down Refinement

Finally, in the top-down refinement step, the abstract nodes pass messages back to their children to update

base topic embeddings with global context (e.g., each v_i may receive an additive message from its abstract parent $\widetilde{v}_k$). In our implementation, we incorporate this by concatenating the parent's representation to the base node before a final linear transformation.

The resulting model forms a hierarchical topic structure: base-level nodes represent fine-grained concepts, while abstract-level nodes capture broader thematic categories. Directional GNN propagation ensures semantic consistency across levels. This design is inspired by hierarchical topic models such as hLDA, which organize topics using Bayesian priors. In contrast, our approach employs deterministic clustering and neural message passing to simultaneously model intra-level causal dependencies and inter-level abstractions. The dual-layer GNN enhances interpretability through structured topic organization and improves predictive performance by enabling statistical sharing among related topics.

3.6 Counterfactual Generation with WGAN and BERT Consistency

A central motivation of our causal topic model is to enable counterfactual reasoning—i.e., answering "what if" questions by generating text under hypothetical interventions. To this end, we leverage the learned causal graph and SCM to guide a text generation module that produces counterfactual documents, while enforcing semantic consistency via a pretrained language model (BERT). Our approach adopts a conditional generative adversarial network, where the generator receives an intervened topic representation and outputs synthetic text, and the discriminator distinguishes between real and generated samples. We employ the Wasserstein GAN variant to enhance training stability and mode coverage (Gulrajani et al., 2017).

3.6.1 Conditioning on Causal Topics

The generator is conditioned on the causal topic vector of a document, which we obtain from the Neural SCM/GNN. For each real document d, we first infer its topic activation vector $z = [z_1, \ldots, z_m]$ using the current model – this can be done by feeding d through the GNN encoder or by direct inference in the SCM. We then sample an intervention on z: for instance, to generate a counterfactual where topic v_k is altered, we set z_k to a new value z'_k (e.g., zero to simulate removing that topic, or a higher value to simulate emphasizing it) while keeping other $z_{i\neq k}$ the same, or also updating descendants of $v_k$via the SCM to reflect causal effects. Denote this intervened topic vector as $z^{cf} = do\left(v_k = z'_k\right)$.

According to our causal model, z^{cf} represents a coherent counterfactual state of the topics (the distribution that would occur if v_k were set to z'_k). The generator G then maps $\left(z^{cf}, \xi\right) \to \widetilde{x}$ where ξ is random noise and $\widetilde{x}$ is a generated text. In practice, G can be implemented as a transformer-based language model or any sequence decoder that accepts a conditioning vector (here z^{cf}may be fed through a projection and used as the initial hidden state or as a prompt). The discriminator D is trained to distinguish real document x from generated $\widetilde{x}$ while G is trained to fool D.

We optimize the standard WGAN objective with gradient penalty,

$$\max_D \min_G \mathbb{E}_{x\sim P_{\text{data}}}[D(x)] - \mathbb{E}_{z,\xi}[D(G(z,\xi))] + \lambda \mathbb{E}\left[(\| \nabla_{\widehat{x}} D(\widehat{x}) \|_2 - 1)^2\right] \quad (10)$$

which guides G to produce outputs whose distribution matches the real text distribution (under various interventions z). Here $\widehat{x}$ denotes points along the line between real and generated samples for the gradient penalty. Crucially, because z is drawn from our causal model (including interventional cases that may not appear in the training data), the generator learns to produce texts for both observational and counterfactual topic combinations.

This approach is analogous to the CausalGAN framework (Kocaoglu et al., 2017), where a generator architecture consistent with a causal graph can output samples from both the true observational and interventional distributions. In our case, the SCM provides z for interventions, and the conditional GAN learns to map those to realistic text, effectively learning $P(\text{text}|\text{topics})$ for both normal and intervened topics. Prior work has demonstrated the feasibility of using WGAN to generate data consistent with a given causal graph, even for interventions not seen in the training set.

3.6.2 BERT-Based Consistency Regularization

While the GAN loss ensures the generated text is realistic, we also want the counterfactual text $\widetilde{x}$ to remain maximally similar to the original text x in all aspects except those affected by the intervention. We introduce a consistency module using BERT to enforce this. Let $\text{Enc}_{\text{BERT}}(x)$be the contextual embedding of the original text, and likewise $\text{Enc}_{\text{BERT}}(\widetilde{x})$ for the generated text. We add a penalty term:

$$L_{\text{cons}} = \| \text{Enc}_{\text{BERT}}(x) - \text{Enc}_{\text{BERT}}(\widetilde{x}) \| \quad (11)$$

which encourages the generated text to lie close to the original in semantic embedding space. In practice, we compute this as the cosine distance between BERT embeddings of x and $\widetilde{x}$ or as a weighted token-level similarity loss. The idea is to preserve the document's main content, modifying only the topical details related to the intervened variables. The BERT consistency loss guides G to make minimal but precise edits. We also explicitly verify that G has indeed effected the desired change (e.g., by checking that keywords for topic $v_k$are reduced or removed in $\widetilde{x}$).

In summary, our counterfactual generation module produces alternative versions of documents by toggling causal factors, using a WGAN to maintain fluency and realism and a BERT-based regularizer to maintain fidelity to the source content.

3.7 Joint Objective and Training Strategy

All components of our model are trained jointly to ensure that causal structure learning, topic modeling, and text generation inform each other. We formulate a multi-term loss that combines the objectives of the structural modules and the generative modules. Specifically, our overall loss $\mathcal{L}_{\text{total}}$ includes:

- A structure loss $\mathcal{L}_{\text{struct}}$ for fitting the causal graph to data (for example, the negative log-likelihood of the observed topic occurrences under the Neural SCM, or an evidence lower bound as in a VAE), plus any L1/L2 regularization on A to encourage sparse, interpretable graphs.
- The DAG constraint penalty $\lambda_{\text{DAG}}|h(A)|$ to enforce acyclicity (or an augmented Lagrangian term as in NOTEARS).
- The GAN loss terms for text generation (the generator and discriminator Wasserstein losses, denoted $\mathcal{L}_G$ and $\mathcal{L}_D$).
- The BERT consistency loss L_{cons}. We weight these components with hyperparameters to balance their influence:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{struct}} + \lambda_{\text{DAG}} h(A) + \lambda_G \mathcal{L}_G \\ + \lambda_D \mathcal{L}_D + \lambda_{\text{cons}} L_{\text{cons}}\end{aligned} \tag{12}$$

Our training strategy is a hybrid of alternating optimization and stage-wise training. Inspired by two-stage approaches in causal generative modeling, we first train the structure and SCM on observational data alone, then train the GAN generator and discriminator on text generation given the learned topics, and finally fine-tune all components together.

In the first stage, we optimize $\mathcal{L}_{\text{struct}}$ with the DAG constraint (using the augmented Lagrangian or a penalty method) to learn A and the neural parameters of f_i (as well as to learn good initial GNN weights for encoding topics). This may be done via variational inference: e.g., treat the topic activations z_i in each document as latent variables and maximize an Evidence Lower Bound(ELBO), or simply by treating inferred topic vectors (from an unsupervised topic model or the LLM extraction frequencies) as training data for a regression model defined by the SCM.

Once a reasonable causal graph and SCM are learned, we proceed to train the GAN. We generate training pairs $\left(z^{cf}, x\right)$ by taking real documents x, inferring their topic vector$z$randomly sampling interventions on z(to get z^{cf}), and using x as the "real" example associated with original $z$versus $G\left(z^{cf}\right)$ as a "fake" example for the intervened case. The discriminator D learns to judge realism, while G learns to produce plausible text for both actual and hypothetical topic conditions. We incorporate the consistency loss with the original text x during G's updates to ensure counterfactuals remain anchored to x.

In the final joint stage, we allow gradients from the text generation loss to also update the earlier modules (SCM and GNN), which can further refine the topic representations to better support fluent generation. In practice, we alternate between an epoch of structure+SCM/GNN updates (minimizing $\mathcal{L}_{\text{struct}} + \lambda_{\text{DAG}} h(A)$ while keeping GAN fixed), and an epoch of GAN updates (minimizing $\mathcal{L}_G + \lambda_{\text{cons}} L_{\text{cons}}$ and maximizing $\mathcal{L}_D$ with A and SCM fixed). This alternating schedule is similar in spirit to expectation-maximization or the two-phase training of CausalGAN – first learn to model the latent causal factors, then learn to generate observable data from those factors.

Finally, we fine-tune end-to-end on the combined objective $\mathcal{L}_{\text{total}}$ (with a small learning rate) to ensure all parts are mutually consistent. Throughout training, we monitor the DAG constraint and gradually increase the penalty coefficient λ_{DAG} to drive $h(A) \to 0$, ensuring the learned graph remains acyclic.

4 EXPERIMENTS AND RESULTS

4.1 Experimental Setup

4.1.1 Dataset and Preprocessing

We train and evaluate CGNTM on the PubMed Lung Cancer Corpus, consisting of approximately 20,000 English-language articles (titles and abstracts) published over the past two decades. The dataset was built by querying PubMed with domain-specific keywords such as "lung cancer" and "non-small cell carcinoma".

For structured causal inputs, we use an LLM with prompt engineering to extract salient keywords from each document. Contextual patterns infer causal relationships, producing triples in the form ⟨cause, relation, effect⟩ to construct document-level causal graphs for the CGNTM pipeline. The corpus is split into 80% training and 20% testing, with 10% of training held out for hyperparameter tuning.

4.1.2 Evaluation Metrics

We evaluate CGNTM on topic quality and causal correctness using five metrics.

Topic Coherence: Normalized Pointwise Mutual Information (NPMI) measures the average normalized pointwise mutual information among the top words within each topic. For topic t, let W_t denote its top-k words. The NPMI score is computed as:

$$\text{NPMI}(t) = \frac{2}{k(k-1)} \sum_{1 \le i < j \le k} \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \quad (13)$$

where $P(w_i)$ and $P(w_i, w_j)$ are estimated from corpus-wide co-occurrence counts. Higher NPMI indicates stronger semantic coherence and aligns better with human judgment.

Topic Diversity (TD): Measures the proportion of unique words across all topics. Let T denote the set of all topics and V_{top} the set of all top-k words across topics. Then:

$$\text{TD} = \frac{|\bigcup_{t \in T} W_t|}{k \cdot |T|} \quad (14)$$

Higher TD implies broader topic coverage and lower redundancy.

Causal Precision (CP): The proportion of inferred causal edges $(i \to j) \in E$ that match a curated biomedical causal knowledge base E^*:

$$\text{CP} = \frac{|E \cap E^*|}{|E|} \quad (15)$$

The CP metric is a standard precision measure in causal discovery literature, directly adapted from the predictive model evaluation metric "Precision" to assess the accuracy of inferred causal edges. Higher CP indicates better alignment with known causal relations.

Reverse Causality Rate (RCR): Measures the fraction of inferred causal edges that contradict known causal directionality:

$$\text{RCR} = \frac{|\{(i \to j) \in E \mid (j \to i) \in E^*\}|}{|E|} \quad (16)$$

A lower RCR suggests more accurate directional inference.

Counterfactual Semantic Alignment (CSA): Assesses whether interventions modify only the targeted causal components while preserving unrelated semantics. Let x be a document, $\tilde{x}$ its counterfactual version, and $\phi(\cdot)$the [CLS] embedding from BERT. Then:

$$\text{CSA}(x, \tilde{x}) = \cos(\phi(x), \phi(\tilde{x})) \quad (17)$$

where cosine similarity is used to measure semantic alignment. Higher CSA reflects more precise and faithful counterfactual generation.

4.1.3 Implementation Details

CGNTM is implemented in PyTorch, using a 2-layer GAT with directional masking for causal graph propagation and a 3-layer MLP for nonlinear SCM dependencies. Hierarchical modeling aggregates base-level topic embeddings into abstract topics. Counterfactual generation employs a WGAN with gradient penalty ($\lambda = 10$) and BERT-based regularization. Optimization uses Adam (learning rate 0.001, batch size 16) for up to 100 epochs with early stopping. Experiments run on a GPU, hyperparameters are tuned on a validation set. Source code is available at https://github.com/Longcchao-Wang/Causal-Topic for reproducibility.

4.2 Quantitative and Qualitative Results

4.2.1 Comparison with Baselines

We evaluate CGNTM against six representative baselines spanning classical, neural, graph-based, and causal paradigms, focusing on models that support probabilistic latent factor modeling for unified topic discovery and causal inference. Recent semantic clustering approaches like BERTopic and Top2Vec

achieve strong coherence via pre-trained embeddings but lack latent variables essential for causal structure learning, hence their exclusion. The baselines are:

- CRNTM(Tang et al., 2024): A supervised causal model learning relations among latent topics and labels via structural equation modeling over a DAG; supervision simulated with synthetic biomedical risk factor labels for fair comparison (Tang et al., 2024).
- LDA(Blei et al., 2003): A classical probabilistic model with Dirichlet priors on document–topic and topic–word distributions.
- ETM(Dieng et al., 2020): A neural topic model projecting words into continuous latent spaces to improve topic coherence.
- NVDM(Miao et al., 2016): A variational autoencoder encoding documents as latent vectors from bag-of-words input.
- GNTM(Shen et al., 2021): A graph-based neural model using document-level word co-occurrence graphs and GNNs.
- GNTM-CK(Zhu et al., 2023): GNTM extended with ConceptNet commonsense knowledge.

These facilitate comparisons across supervised vs. unsupervised, causal vs. correlational, and knowledge-enhanced vs. data-driven paradigms. Table 1 summarizes performance across NPMI, TD, CP, RCR, and CSA, showing CGNTM outperforms all in coherence, diversity, and causal alignment.

For topic quality, CGNTM achieves the highest NPMI (0.30), surpassing supervised CRNTM (0.29) and unsupervised baselines like GNTM-CK (0.26) and ETM (0.24), indicating superior coherence from biomedical priors and causal constraints. It also leads in TD (0.82), reflecting broader coverage with minimal redundancy via hierarchical modeling and synonym-aware clustering.

In causal accuracy, only CRNTM and CGNTM produce explicit graphs; CGNTM's CP (0.70) nears CRNTM (0.80), uncovering meaningful biomedical causalities without supervision, while its low RCR (0.10) confirms reliable directionality close to CRNTM (0.07). Non-causal models are N/A for CP/RCR.

For CSA, CGNTM's 0.88 surpasses all baselines, ensuring counterfactuals remain semantically consistent except for targeted interventions, unlike CRNTM's 0.80 due to lacking explicit counterfactual training. Other models lack intervention support, rendering CSA inapplicable.

Table 1: Comparison of CGNTM with baseline models.

Model	NPMI	TD	CP	RCR	CSA
LDA	0.18	0.81	0.51	N/A	N/A
NVDM	0.22	0.72	0.53	N/A	N/A
ETM	0.24	0.73	0.56	N/A	N/A
GNTM	0.25	0.76	0.58	N/A	N/A
GNTM-CK	0.26	0.77	0.63	N/A	N/A
CRNTM	0.29	0.80	0.69	0.07	0.80
CGNTM (ours)	0.30	0.82	0.70	0.10	0.88

4.2.2 Ablation Study

To assess the contribution of individual components within CGNTM, we conduct an ablation study with four modified variants (denoted as "w/o" for "without"): w/o LLM Extraction, w/o Neural SCM, w/o WGAN + Consistency, and w/o Hierarchy, as summarized in Table 2.

a) w/o LLM Extraction: Replaces LLM-based keyword and causal triple extraction with co-occurrence-based graphs (e.g., PMI edges). Performance drops in CP and CSA validate the importance of knowledge-guided structure.

b) w/o Neural SCM: Replaces the nonlinear Structural Causal Model with a linear or identity mapping, disabling deep causal propagation. NPMI and CP decline, highlighting the benefit of modeling nonlinear causal effects.

c) w/o WGAN + Consistency: Removes the counterfactual generation and semantic consistency loss. While core topic metrics remain stable, CSA significantly drops, confirming the WGAN's role in ensuring targeted and semantically aligned interventions.

d) w/o Hierarchy: Flattens the topic structure by removing macro-micro topic separation. TD decreases due to more redundancy, and NPMI also slightly declines, suggesting that hierarchical modeling improves topic specialization.

Table 2: Ablation results for CGNTM.

Model	NPMI	TD	CP	RCR	CSA
Full CGNTM	0.30	0.82	0.70	0.10	0.88
(–) LLM Extraction	0.27	0.80	0.61	0.15	0.79
(–) Neural SCM	0.28	0.81	0.64	0.13	0.83
(–) WGAN +Consistency	0.29	0.81	0.67	0.11	0.76
(–) Hierarchical Structure	0.28	0.76	0.66	0.12	0.84

4.3 Hyperparameter Sensitivity

We evaluate the robustness of CGNTM with respect to two key hyperparameters: the number of topics (K) and the knowledge weight (λ), which controls the influence of the concept graph.

Number of Topics (K): We varied K from 20 to 100 and observed its impact on NPMI and CSA. Topic coherence (NPMI) improves as K increases, peaking around $K = 50$, then plateaus or slightly declines as topics become too fine-grained (e.g., 0.30 at $K = 50$ vs. 0.29 at $K = 100$). Topic diversity grows with K, but with diminishing returns after 50. In terms of causal metrics, CSA peaks in the range of $K = 50$-60, balancing coherence and coverage. Too few topics ($K = 20$) yield broad, less specific topics (CSA $\approx 45\%$), while too many ($K = 100$) introduce redundancy and fragment topic quality.

Knowledge Weight (λ): We tested λ in the range $[0, 1.0]$. At $\lambda = 0$ (no concept supervision), CP and CSA drop substantially, as expected. Increasing λ to 0.5 steadily improves causal metrics, with CSA rising from $\sim$45% to $\sim$58%. However, too high a weight ($\lambda = 1.0$) slightly reduces coherence (NPMI ≈ 0.285), as the model may overfit to concept connections. We found $\lambda = 0.5$–0.7 provides the best trade-off, and used $\lambda = 0.6$ as default.

Overall, CGNTM shows stable performance across a wide hyperparameter range. We recommend $K \approx 50$ and λ in $[0.5, 0.7]$ for corpora of similar size and domain complexity. These results confirm that CGNTM's gains are not contingent on narrow hyperparameter settings, but stem from the model's design.

4.4 Summary and Discussion

The results underscore CGNTM's strengths as the first unsupervised topic model integrating LLM-based extraction with neural causal modeling to uncover interpretable causal relations. Relying solely on unlabeled data, CGNTM matches supervised CRNTM in performance while discovering novel causalities beyond label structures and organizing topics hierarchically—a capability CRNTM lacks.

Compared to unsupervised models like NVDM, CGNTM excels in coherence and diversity, constructing directed graphs for causal reasoning. For example, it infers directionality (e.g., EGFR mutation $\rightarrow$ drug resistance) where traditional models merely co-cluster terms, enabling counterfactual simulation and hypothesis generation.

Relative to NVDM, CGNTM enhances quality via causal regularization, avoiding posterior collapse. Unlike BERTopic's embedding clustering without causality, CGNTM's generative framework (SCM, directional GNN, causal priors) captures both semantic and causal structures.

In summary, CGNTM bridges contextual topic modeling and causal discovery, advancing unsupervised methods with descriptive and explanatory insights aligned to domain knowledge.

5 CONCLUSION

We introduce CGNTM, the first unsupervised causal topic model merging LLM knowledge extraction with neural causal inference. It discovers interpretable topics and domain-reflective causal graphs without labels, achieving competitive coherence and diversity while enabling counterfactual reasoning through structured SCM and GNN design. This supports explanatory modeling, inferring relations like "EGFR mutation $\rightarrow$ drug resistance" from text-valuable for biomedicine and social sciences.

Limitations include dependence on LLM triple quality (errors impact inference), computational intensity (BERT embeddings, GNN propagation, adversarial training), and causal evaluation challenges from limited ground-truth.

Future work involves multilingual/cross-domain applications, semi-supervised signals (e.g., seed causal edges), and structured knowledge bases for graph constraints.

Ultimately, CGNTM advances topic modeling by embedding causal discovery unsupervised, fostering automated hypothesis generation beyond "what" to "why".

ACKNOWLEDGEMENTS

This research was funded by the Innovation Fund for Medical Sciences of Chinese Academy of Medical Sciences grant number 2021-I2M-1-033.

REFERENCES

Behnam, A. and Wang, B. (2024). Graph neural network causal explanation via neural causal models. In *European Conference on Computer Vision*, pages 410–427. Springer Nature Switzerland.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Chen, L., Ban, T., Wang, X., Lyu, D., and Chen, H. (2023). Mitigating prior errors in causal structure learning:

Towards llm driven prior knowledge. arXiv preprint arXiv:2306.07032.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Gao, H., Yao, C., Li, J., Si, L., Jin, Y., Wu, F., and Liu, H. (2024). Rethinking causal relationships learning in graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12145–12154.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30.

Kaushik, D., Hovy, E., and Lipton, Z. C. (2019). Learning the difference that makes a difference with counterfactually-augmented data. arXiv preprint arXiv:1909.12434.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. (2017). Causalgan: Learning causal implicit generative models with adversarial training. arXiv preprint arXiv:1709.02023.

Lagemann, K., Lagemann, C., Taschler, B., and Mukherjee, S. (2023). Deep learning of causal structures in high dimensions under data limitations. *Nature Machine Intelligence*, 5(11):1306–1316.

Liu, Z., Grau-Bove, J., and Orr, S. A. (2022). Bert-flow-vae: a weakly-supervised model for multi-label text classification. arXiv preprint arXiv:2210.15225.

Miao, Y., Yu, L., and Blunsom, P. (2016). Neural variational inference for text processing. In *International Conference on Machine Learning*, pages 1727–1736. PMLR.

Morstatter, F. and Liu, H. (2018). In search of coherence and consensus: measuring the interpretability of statistical topics. *Journal of Machine Learning Research*, 18(169):1–32.

Panwar, M., Shailabh, S., Aggarwal, M., and Krishnamurthy, B. (2020). Tan-ntm: Topic attention networks for neural topic modeling. arXiv preprint arXiv:2012.01524.

Park, S. and Kim, J. (2023). Dag-gcn: directed acyclic causal graph discovery from real world data using graph convolutional networks. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 318–319. IEEE.

Pawlowski, N., de Castro, D. C., and Glocker, B. (2020). Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, pages 857–869.

Prostmaier, B., Vávra, J., Grün, B., and Hofmarcher, P. (2025). Seeded poisson factorization: Leveraging domain knowledge to fit topic models. arXiv preprint arXiv:2503.02741.

Rana, M., Hacioglu, K., Gopalan, S., and Boothalingam, M. (2024). Zero-shot slot filling in the age of llms for dialogue systems. arXiv preprint arXiv:2411.18980.

Shen, D., Qin, C., Wang, C., Dong, Z., Zhu, H., and Xiong, H. (2021). Topic modeling revisited: A document graph-based neural network perspective. In *Advances in Neural Information Processing Systems*, volume 34, pages 14681–14693.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488.

Tang, Y. K., Huang, H., Shi, X., and Mao, X. L. (2024). Beyond labels and topics: Discovering causal relationships in neural topic modeling. In *Proceedings of the ACM Web Conference 2024*, pages 4460–4469.

Venugopalan, M. and Gupta, D. (2022). An enhanced guided lda model augmented with bert based semantic strength for aspect term extraction in sentiment analysis. *Knowledge-based Systems*, 246:108668.

Wang, B., Li, J., Chang, H., Zhang, K., and Tsung, F. (2025). Heterophilic graph neural networks optimization with causal message-passing. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 829–837.

Wu, Y., McConnell, L., and Iriondo, C. (2024). Counterfactual generative modeling with variational causal inference. arXiv preprint arXiv:2410.12730.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2021). Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602.

Yang, Y., Nafea, M. S., Ghassami, A., and Kiyavash, N. (2022). Causal discovery in linear structural causal models with deterministic relations. In *Conference on Causal Learning and Reasoning*, pages 944–993. PMLR.

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019). Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR.

Zečević, M., Dhami, D. S., Veličković, P., and Kersting, K. (2021). Relating graph neural networks to structural causal models. arXiv preprint arXiv:2109.04173.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31.

Zhu, B., Cai, Y., and Ren, H. (2023). Graph neural topic model with commonsense knowledge. *Information Processing & Management*, 60(2):103215.

Zhu, Q., Feng, Z., and Li, X. (2018). Graphbtm: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4663–4672.