# Disentangled Graph-Enhanced Large Language Models for Fair Learning

**Zhipeng Yin**[1] , **Zichong Wang**[1] , **Zhong Chen**[2] ,
**Jack Yang**[3] , **Xin Ning**[1] and **Wenbin Zhang**[1*]
[1]Florida International University, FL, USA
[2]Southern Illinois University Carbondale, IL, USA
[3]University of Wollongong, NSW, Australia

## Abstract

Large Language Models (LLMs) achieve strong performance in many applications but remain limited in handling graph-structured data due to their reliance on textual context. Recent approaches integrate Graph Neural Networks (GNNs) to enhance structural modeling, yet they largely overlook fairness, leaving models vulnerable to bias amplification across graph and text modalities. To address this issue, we propose *FairGEnt*, a disentangled graph-enhanced large language model for fair graph learning. FairGEnt separates sensitive-related and sensitive-invariant factors in both graph and textual representations to mitigate bias while preserving task-relevant information, and further aligns the two modalities through a fairness-aware integration module. In addition, FairGEnt incorporates fair graph-enhanced instruction tuning to improve LLM understanding of complex graph structures. Experiments on multiple benchmark datasets demonstrate that FairGEnt consistently outperforms existing methods in both fairness and predictive performance.

## 1 Introduction

Large Language Models (LLMs) have achieved significant advancements in various natural language processing tasks, such as text classification [Zhang *et al.*, 2025b], sentiment analysis [Zhang *et al.*, 2024b], language translation [Gong *et al.*, 2024], and dialog systems [Yi *et al.*, 2024]. These successes can be primarily attributed to the Transformer architecture, which effectively models long-range dependencies in sequential data, combined with extensive training on massive textual datasets [Mienye, 2025]. Transformer-based models are highly dependent on attention mechanisms to capture semantic and contextual nuances from textual input, enabling them to generalize effectively to various downstream tasks [Wang *et al.*, 2025c]. However, despite their proven effectiveness on textual information, LLMs face notable challenges when applied to graph-structured data [Li *et al.*, 2025]. Unlike linear and sequential text, graphs represent non-linear relational structures characterized by nodes and edges, lacking inherent ordering and spatial locality [Osman and Barukub, 2020]. This fundamental difference impedes the direct application of LLMs, as their architectures are optimized for textual rather than structural dependencies [Osman and Barukub, 2020; Wang *et al.*, 2025d]. Addressing this limitation is crucial, as graph data play an increasingly important role in numerous real-world domains, such as social network analysis [Campbell *et al.*, 2013], drug discovery [Bonner *et al.*, 2022], recommendation systems [Yang *et al.*, 2025], and fraud detection [Pourhabibi *et al.*, 2020]. In these scenarios, effectively using the rich structural relationships embedded in graph data is essential for making accurate, reliable, and context-aware predictions. Therefore, addressing the current limitations of LLMs in handling graph-structured data represents an important and timely research challenge.

Building upon this motivation, considerable research has sought to bridge the gap between graph data and LLMs, exploring innovative strategies for effective integration. For instance, GaLM [Xie *et al.*, 2023] leverages graph-aware language modeling to jointly incorporate structural graph information and textual semantics, enabling language models to better capture contextual relationships within graph-structured data. Similarly, GraphTranslator [Zhang *et al.*, 2024a] introduces a translation-based framework that converts graph-structured inputs into natural language prompts through structural linearization, allowing LLMs to process and reason over structured graph data. Despite these important contributions, existing approaches largely overlook critical fairness considerations, inadvertently perpetuating biases during the integration of graph data with textual representations. Specifically, beyond biases already present within the graph data itself, during alignment between graph embeddings and their corresponding textual descriptions, training objectives may unintentionally amplify biases by reinforcing correlations between sensitive attributes and task-relevant semantic information. Additionally, another significant source of bias emerges in the serialization step, whereby complex graph structures are transformed into linear token sequences suitable for LLM input. This serialization process can inadvertently distort the underlying connectivity patterns, disproportionately emphasizing nodes or edges associated with sensitive attributes, thereby introducing structural biases into downstream predictions.

*Corresponding author.

Addressing these fairness-related limitations is essential, as enabling LLMs to process graph-structured data in a fair and unbiased manner is of significant practical importance, yet there are still several key challenges to achieving fair graph-enhanced LLMs: **i) In graph-enhanced LLM pipelines, sensitive attributes and task-relevant relational features are frequently co-encoded during graph-to-text representation learning.** This entanglement makes it difficult to prevent protected signals from being embedded into language representations, which can cause downstream LLM predictions to depend on sensitive information. As a result, reliable fairness control and bias mitigation become substantially more challenging. **ii) Biases can arise from multiple interconnected sources, notably during the alignment of structural and textual modalities.** When graph representations are synchronized with textual descriptions, biased correlations inherent in either domain can propagate and reinforce each other, amplifying stereotypical associations. **iii) A substantial modality gap exists between graph-based models and language models due to fundamental differences in data representation and processing mechanisms.** LLMs operate on token sequences and are optimized for natural language understanding and generation, rather than directly modeling relational structures. Bridging this gap while suppressing bias transmission requires carefully designed mechanisms that preserve structural semantics without introducing or amplifying unfair correlations during graph-to-text conversion. These interconnected challenges underscore the necessity of developing fairness-aware approaches explicitly tailored for integrating graph data with LLMs.

To tackle the aforementioned challenges, we propose *Disentangled Graph-Enhanced Large Language Model* (FairGEnt), a novel disentangled graph-enhanced large language model explicitly designed to promote fairness in LLM-based graph learning. *To the best of our knowledge, FairGEnt is the first work that explicitly addresses fairness in graph-enhanced large language models by jointly disentangling and aligning sensitive-related structural information and task-relevant representations, thereby effectively mitigating bias transmission across modalities.* Specifically, FairGEnt first disentangles sensitive-related and task-relevant structural factors in both graph and textual representations, effectively mitigating bias caused by correlated sensitive attributes. It then employs a fairness-aware alignment module to achieve unbiased integration between the disentangled graph embeddings and textual semantics. In addition, FairGEnt incorporates a fair graph-enhanced instruction-tuning strategy, enabling large language models to better understand complex structural contexts while preventing bias amplification. The key contributions of this work are summarized as follows:

- We study a new fairness problem in graph-enhanced LLMs: how to mitigate the implicit entanglement between sensitive attributes and task-relevant relational information during graph-to-text representation learning, paving the way for achieving fair node prediction with LLM-based graph learning.
- We propose *FairGEnt*, a disentangled graph-enhanced LLM explicitly designed to promote fairness in graph-LLM integration. It systematically separates sensitive related and invariant attributes, leverages fairness-aware alignment and instruction-tuning strategies, enabling accurate and equitable node classification.
- Through extensive evaluations in multiple real-world datasets, we demonstrate that *FairGEnt* significantly improves both fairness and node classification accuracy compared to existing state-of-the-art methods.

## 2 Related Work

### 2.1 Graph-enhanced LLMs

LLMs have exhibited impressive performance in diverse natural language processing tasks [Lopes *et al.*, 2020; Azher *et al.*, 2024], yet their ability to directly process and reason over graph data remains inherently limited due to their sequential nature. Recent research has sought to bridge this gap by developing graph-enhanced frameworks that integrate graph data into language modeling. For example, GaLM [Xie *et al.*, 2023] jointly models graph structural information and textual semantics through graph-aware language modeling, enabling language models to better capture contextual dependencies within graph data. Similarly, GraphLLM [Chai *et al.*, 2025] introduces a specialized graph transformer architecture, incorporating learnable structural queries and positional encodings to encode graph representations that guide prefix-tuning strategies within LLMs. HiGPT [Tang *et al.*, 2024] advances hierarchical graph encoding, generating structure-aware prompts through multi-granularity encoding to improve language model reasoning over complex graphs. Moreover, GraphTranslator [Zhang *et al.*, 2024a] employs contrastive pre-training to facilitate semantic alignment between graph structures and textual contexts, significantly enriching LLM representations. Although these approaches extend the capabilities of LLMs to structured data, they largely neglect fairness considerations, leading to biases that exacerbate discrimination across sensitive attributes.

### 2.2 Fairness in LLMs

Fairness has gained considerable attention in natural language processing, as pretrained LLMs inherently absorb and replicate societal biases from large-scale textual datasets [Wang *et al.*, 2025a]. These biases manifest prominently through stereotypical associations within contextual embeddings, uneven outcomes in masked language modeling tasks [Radaideh *et al.*, 2025], and skewed reasoning patterns observed during chain-of-thought prompting [Wei *et al.*, 2022]. To mitigate such biases, existing strategies include adversarial training to enforce demographic invariance [Chinta *et al.*, 2024], counterfactual data augmentation for robust representation learning [Zmigrod *et al.*, 2019], and carefully designed fairness-oriented prompting techniques [Chu *et al.*, 2024]. However, these fairness-enhancing approaches primarily focus on textual information alone, neglecting structural biases embedded within graph-structured data and thus limiting their effectiveness in multi-modal or cross-domain scenarios that integrate graph and textual data.

Different from these works, our approach explicitly addresses the fairness challenges that arise when training LLMs

on graph data. Specifically, we propose a fairness-aware graph learning model to guide LLM training by disentangling sensitive contextual information from task-related features, thus enabling the model to make fair predictions without compromising task performance. Through this disentanglement, our method effectively reduces biases originating from sensitive attributes embedded within the graph, providing a direct solution to critical limitations identified in previous research.

## 3 Notations

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ represent an undirected attributed graph, where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ denotes the node set with $|\mathcal{V}| = n$, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of edges. We also let $\Phi^G(\cdot)$ and $\Phi^L(\cdot)$ represent the graph encoder and the language model encoder, respectively, mapping input data into corresponding latent representations. Specifically, we define two types of input data: $\mathbf{X} \in \mathbb{R}^{n \times d}$ denotes the node feature data and $P$ denotes the textual data provided to the LLM. The resulting embedding representations are denoted by $\mathbf{H}^G$ for the graph model and $\mathbf{H}^L$ for the LLM. Each node $v_i \in \mathcal{V}$ is associated with a binary sensitive attribute $s_i \in \{0, 1\}$, collectively denoted as $\mathbf{S} \in \{0, 1\}^{n \times 1}$. The nodes with $s_i = 0$ constitute the deprived group $\mathbf{S}_d$ (*e.g.*, female), whereas nodes with $s_i = 1$ form the favored group $\mathbf{S}_f$ (*e.g.*, male). The adjacency matrix is represented by $\mathbf{A} \in \{0, 1\}^{n \times n}$, where $\mathbf{A}_{ij} = 1$ indicates an edge between the nodes $v_i$ and $v_j$, and 0 otherwise. Furthermore, we use $y_i \in \{0, 1\}$ to indicate the binary ground-truth label for each node $v_i$, and $\hat{y}_i^{\mathrm{G}}$ to denote its predicted label from the graph model. Lastly, the predictions obtained from the language model are denoted as $\hat{y}_i^{\mathrm{L}}$.

## 4 Methodology

### 4.1 The Proposed Model - FairGEnt

In this section, we introduce the novel graph-enhanced LLM, *FairGEnt*, which systematically ensures fairness-aware representation learning and enhances downstream reasoning performance by disentangling sensitive attributes from task-relevant information. The *FairGEnt* framework consists of three core components: i) Fairness-guided Graph-enhanced Encoder Module, ii) Fair Graph-enhanced Alignment Module, and iii) Fair Graph-enhanced Instruction Tuning strategy. Specifically, we first propose the Fairness-guided graph-enhanced Encoder, which explicitly separates sensitive attributes from task-relevant attributes within both graph and textual data, producing sensitive-invariant embeddings to suppress potential bias leakage. We then introduce the fair graph-enhanced alignment module, which ensures fairness-aware integration between invariant embeddings and LLM representations. Finally, the Fair Graph-enhanced Instruction Tuning strategy utilizes aligned invariant representations to generate LLM-compatible graph tokens, further enhancing fairness-aware predictive performance through self-supervised and task instruction tuning.

### 4.2 Fairness-guided disentangled Encoder Module

We begin by introducing FairGEnt's disentanglement mechanism, which employs a fairness-guided encoder composed of two key components: **(i) graph disentanglement**, which separates sensitive-related information from sensitive-invariant structural representations within graph embeddings; and **(ii) textual graph disentanglement**, which isolates sensitive semantic content from task-relevant textual representations obtained from an LLM. In the subsequent discussion, we elaborate on the graph disentanglement component in detail.

**Graph disentanglement.** To achieve graph structural disentanglement, we introduce a Disentangled Graph Variational Autoencoder (DG-VAE) that explicitly partitions the graph structure into two separate latent variables: sensitive-related ($\mathbf{E}^{\mathrm{sen}}$) and sensitive-invariant ($\mathbf{E}^{\mathrm{inv}}$). Leveraging variational Bayesian inference, our approach captures sensitive and non-sensitive information from the input graph by modeling $\mathbf{E}^{\mathrm{sen}}$ and $\mathbf{E}^{\mathrm{inv}}$ as latent variables within a Bayesian network. Specifically, our framework comprises two inference networks, parameterized by $\phi$ and $\theta$, approximating posterior distributions defined as: $q_\phi(\mathbf{E}^{\mathrm{sen}} \mid \mathbf{A}, \mathbf{X}, \mathbf{S})$ and $q_\theta(\mathbf{E}^{\mathrm{inv}} \mid \mathbf{A}, \mathbf{X})$, respectively. The sensitive-related encoder explicitly utilizes sensitive attributes ($\mathbf{S}$) to capture structural features directly linked to sensitive factors, whereas the invariant encoder intentionally excludes $\mathbf{S}$ to ensure the resulting latent representations are insensitive to sensitive attributes. To further strengthen the disentanglement beyond the explicit disentanglement architecture, we introduce an additional independence constraint by minimizing the Hirschfeld–Gebelein–Rényi (HGR) maximal correlation between $\mathbf{E}^{\mathrm{sen}}$ and $\mathbf{E}^{\mathrm{inv}}$, as detail below:

$$\mathrm{HGR}\big(\mathbf{E}^{\mathrm{sen}}, \mathbf{E}^{\mathrm{inv}}\big) = \sup_{f_{\mathrm{sen}}^G,\, f_{\mathrm{inv}}^G} \frac{\mathbb{E}\big[f_{\mathrm{sen}}^G(\mathbf{E}^{\mathrm{sen}}) f_{\mathrm{inv}}^G(\mathbf{E}^{\mathrm{inv}})\big]}{\sqrt{\mathbb{E}[(f_{\mathrm{sen}}^G)^2(\mathbf{E}^{\mathrm{sen}})]\,\mathbb{E}\big[(f_{\mathrm{inv}}^G)^2(\mathbf{E}^{\mathrm{inv}})\big]}} \tag{1}$$

where $f_{sen}^G$ and $f_{inv}^G$ are learnable functions with zero mean and unit variance. This term is optimized in a min–max fashion: adversarial networks are trained to maximize the correlation estimate, while the encoders are updated to minimize it, thereby encouraging $\mathbf{E}^{\mathrm{sen}}$ and $\mathbf{E}^{\mathrm{inv}}$ to be maximally independent.

Building on this, the final loss function of the structural graph disentanglement component is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{graph}} = \mathbb{E}_{q_{\phi,\theta}}\big[&\log p_\psi\big(\mathbf{A} \mid \mathbf{E}^{\mathrm{sen}}, \mathbf{E}^{\mathrm{inv}}\big) \\ &+ \log p_\omega\big(\mathbf{X}_S \mid \mathbf{A}, \mathbf{E}^{\mathrm{sen}}, \mathbf{E}^{\mathrm{inv}}\big)\big] \\ &+ \log p_\omega\big(\mathbf{X}_{\overline{S}} \mid \mathbf{A}, \mathbf{E}^{\mathrm{sen}}, \mathbf{E}^{\mathrm{inv}}\big) \\ &- \mathrm{KL}(q_\phi(\mathbf{E}^{\mathrm{sen}} \mid \mathbf{A}, \mathbf{X}, \mathbf{S}) \parallel p(\mathbf{E}^{\mathrm{sen}})) \\ &- \mathrm{KL}\big(q_\theta\big(\mathbf{E}^{\mathrm{inv}} \mid \mathbf{A}, \mathbf{X}\big) \parallel p\big(\mathbf{E}^{\mathrm{inv}}\big)\big) \\ &+ \lambda_{HGR}\,\mathrm{HGR}\big(\mathbf{E}^{\mathrm{sen}}, \mathbf{E}^{\mathrm{inv}}\big)\end{aligned} \tag{2}$$

where the first three terms constitute the Evidence Lower Bound (ELBO), maximizing the likelihood of reconstructing the graph structure and node attributes while regularizing the variational posterior distributions toward their priors. $\lambda_{HGR}$ is a hyperparameter balancing the strength of the independence constraint. In summary, this module explicitly disen-

tangles graph structures into sensitive-related and sensitive-invariant latent representations. The disentanglement is further enhanced by applying the HGR-based independence constraint, effectively minimizing correlation between these representations. By clearly isolating sensitive structural information, this module provides explicit guidance for training LLMs on graph-based tasks.

**Textual graph disentanglement.** In our previous disentanglement stage, we successfully separated sensitive attributes from sensitive-invariant information within graph embeddings. Nevertheless, accurate alignment between graph structures and textual semantics necessitates a similar disentanglement within textual embeddings generated by LLMs. To achieve this, we introduce a dedicated textual graph disentanglement approach, explicitly decomposing textual embeddings into sensitive and sensitive-invariant components, guided by the previously disentangled graph representations, enhancing fairness-aware alignment.

Given a graph $\mathcal{G}$, we sample local subgraphs and serialize them into structured textual prompts using *GraphML* [Brandes *et al.*, 2013], preserving node features and relational structures in structured textual form. Each textual prompt $\mathcal{P}_n$ is fed into a pretrained LLM encoder $\Phi_L(\cdot)$ (*i.e.*, BERT [Koroteev, 2021]) to obtain subgraph-level embeddings: $\mathbf{H}_n^L = \Phi_L(\mathcal{P}_n)$. However, embeddings from LLM encoders often entangle task-relevant information with sensitive attribute information, which can lead to bias.

To control and disentangle these factors, we design two lightweight projection networks to decompose $\mathbf{H}_n^L$ into complementary subspaces. Specifically, we introduce two lightweight multilayer perceptron projection heads, $f_{\text{sen}}^L(\cdot)$ and $f_{\text{inv}}^L(\cdot)$, to decompose the latent representation as follows:

$$\mathbf{z}_n^{\text{sen}} = f_{\text{sen}}^L(\mathbf{H}_n^L), \quad \mathbf{z}_n^{\text{inv}} = f_{\text{inv}}^L(\mathbf{H}_n^L) \tag{3}$$

The embedding $\mathbf{z}_n^{\text{sen}}$ is intended to capture sensitive-related factors, while $\mathbf{z}_n^{\text{inv}}$ encodes sensitive-invariant and task-relevant information that is invariant to sensitive attributes.

To ensure effective disentanglement, we impose an explicit independence constraint between the sensitive-related and sensitive-invariant textual embeddings, with the goal of reducing statistical dependence and preventing bias leakage during graph–text alignment. To limit information leakage between these two textual subspaces, we introduce a distance covariance-based regularization term as a principled dependence measure. Specifically, given a mini-batch of $B$ samples, each sample $n$ is associated with a pair of textual embeddings $(\mathbf{z}_n^{\text{sen}}, \mathbf{z}_n^{\text{inv}})$, where $\mathbf{z}_n^{\text{sen}}$ captures sensitive-related information and $\mathbf{z}_n^{\text{inv}}$ represents sensitive-invariant content. We collect these embeddings as $\mathbf{Z}_{\text{sen}} = \{\mathbf{z}_n^{\text{sen}}\}_{n=1}^B$ and $\mathbf{Z}_{\text{inv}} = \{\mathbf{z}_n^{\text{inv}}\}_{n=1}^B$, and measure their dependence using distance correlation, defined as:

$$\begin{aligned}\mathcal{L}_{\text{textual}} &= \mathrm{dCov}(\mathbf{Z}_{\text{sen}}, \mathbf{Z}_{\text{inv}}) \\ &= \frac{\mathrm{dCov}(\mathbf{Z}_{\text{sen}}, \mathbf{Z}_{\text{inv}})}{\sqrt{\mathrm{dCov}(\mathbf{Z}_{\text{sen}}, \mathbf{Z}_{\text{sen}})\,\mathrm{dCov}(\mathbf{Z}_{\text{inv}}, \mathbf{Z}_{\text{inv}})}}\end{aligned} \tag{4}$$

where $\mathrm{dCov}(\cdot,\cdot)$ denotes the distance covariance computed from pairwise Euclidean distance matrices after double-centering operations. Minimizing $\mathcal{L}_{\text{textual}}$ ensures statistical independence between sensitive-related and invariant textual representations, effectively suppressing both both linear and non-linear dependencies.

### 4.3 Fair Graph-enhanced Alignment Module

Building upon the disentangled embeddings from the previous step, aligning with invariant attributes is critical to prevent downstream tasks from inheriting biases related to sensitive attributes such as gender or ethnicity, thus promoting fairness. We next propose the *Fair Graph-enhanced Alignment Module*, designed explicitly to align only sensitive-invariant attributes across textual graph embeddings and raw graph representations. Furthermore, invariant attributes encapsulate structural and functional knowledge less susceptible to domain-specific biases, enhancing the robustness and generalizability of learned representations. This alignment is implemented through: (1) lightweight projection into a shared latent space and (2) leveraging neighborhood contexts for enhanced structural coherence.

Given node-level textual graph invariant embeddings $\mathbf{z}_n^{\text{inv}} \in \mathbb{R}^{d_z}$ and graph data sensitive-invariant representations $\mathbf{E}_n^{\text{inv}} \in \mathbb{R}^{d_e}$, we project both into a common alignment space to facilitate cross-modal matching. Formally, we employ lightweight linear transformations defined as follows:

$$\tilde{\mathbf{z}}_n = \frac{\phi_Z(\mathbf{z}_n^{\text{inv}})}{|\phi_Z(\mathbf{z}_n^{\text{inv}})|_2}, \quad \tilde{\mathbf{e}}_n = \frac{\phi_E(\mathbf{E}_n^{\text{inv}})}{|\phi_E(\mathbf{E}_n^{\text{inv}})|_2} \tag{5}$$

where $\phi_Z(\cdot)$ and $\phi_E(\cdot)$ are lightweight linear projection layers mapping inputs into a shared latent space of dimensionality $d$.

After projecting the sensitive-invariant attributes into a shared latent space, a remaining challenge is the potential degradation of alignment quality when relying exclusively on isolated node-level information. To address this issue, we enhance the representation of each node by aggregating information from its immediate neighborhood. This graph-enhanced contextual representation explicitly incorporates local graph topology, yielding embeddings that are structurally coherent and more informative for alignment. We define the neighborhood-enhanced textual representation of node $n$ as:

$$\tilde{\mathbf{z}}_n^{\mathcal{N}} = \frac{1}{|\mathcal{N}(n)|} \sum j \in \mathcal{N}(n) \tilde{\mathbf{z}}_j \tag{6}$$

where $\mathcal{N}(n)$ is the set of neighboring nodes derived from the adjacency matrix of the raw graph. This contextualization ensures that our alignment is robust against local perturbations and captures structural coherence effectively.

To effectively align sensitive-invariant embeddings across modalities, we adopt a multi-component contrastive learning framework. We design three distinct yet complementary similarity measures to facilitate alignment:

*Node-level alignment* captures the immediate semantic alignment between node representations:

$$\mathbf{M}_{\text{node}} = \tilde{\mathbf{E}}\tilde{\mathbf{Z}}^\top, \quad M_{\text{node}}^{(i,j)} = \frac{\tilde{\mathbf{e}}_i \cdot \tilde{\mathbf{z}}_j}{\tau} \tag{7}$$

where $\tau$ is a temperature hyperparameter controlling the distribution sharpness.

*Node-to-neighborhood alignment* explicitly incorporates contextual information, aligning node embeddings with aggregated neighborhood representations from the opposing modality:

$$\mathbf{M}_{\text{ctx}} = \tilde{\mathbf{E}}(\tilde{\mathbf{Z}}^{\mathcal{N}})^{\top}, \quad M_{\text{ctx}}^{(i,j)} = \frac{\tilde{\mathbf{e}}_i \cdot \tilde{\mathbf{z}}_j^{\mathcal{N}}}{\tau} \tag{8}$$

*Intra-modal textual consistency alignment* ensures internal consistency within textual embeddings and their neighborhood contexts, stabilizing embedding distributions:

$$\mathbf{M}_{\text{intra}} = \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^{\mathcal{N}})^{\top}, \quad M_{\text{intra}}^{(i,j)} = \frac{\tilde{\mathbf{z}}_i \cdot \tilde{\mathbf{z}}_j^{\mathcal{N}}}{\tau} \tag{9}$$

We unify these similarity matrices into a joint alignment loss:

$$\begin{aligned}\mathcal{L}_{\text{align}}^{\text{inv}} = \lambda_{\text{node}}\big[\text{CE}(\mathbf{S}_{\text{node}}, \mathbf{y}) + \text{CE}(\mathbf{S}_{\text{node}}^{\top}, \mathbf{y})\big] \\ + \lambda_{\text{ctx}}\big[\text{CE}(\mathbf{S}_{\text{ctx}}, \mathbf{y}) + \text{CE}(\mathbf{S}_{\text{ctx}}^{\top}, \mathbf{y})\big] \\ + \lambda_{\text{intra}}\big[\text{CE}(\mathbf{S}_{\text{intra}}, \mathbf{y}) + \text{CE}(\mathbf{S}_{\text{intra}}^{\top}, \mathbf{y})\big]\end{aligned} \tag{10}$$

where the target vector $\mathbf{y} = (0, 1, \ldots, B-1)$ indicates correct node-to-node correspondences within each batch, and $\lambda$ balances the influence of each component.

To optimize FairGEnt in a unified manner, we combine the alignment loss with the disentanglement losses from both textual and structural disentanglement modules, forming the overall pre-training objective. The total loss function as follow:

$$\mathcal{L}_{\text{pretraining}} = \beta\mathcal{L}_{\text{align}}^{\text{inv}} + \alpha(\mathcal{L}_{textual} + \mathcal{L}_{graph}) \tag{11}$$

where $\alpha$ modulates the strength of disentanglement and $\beta$ controls the alignment intensity.

### 4.4 Fair Graph-enhanced Instruction Tuning

Following the fair graph-enhanced alignment established in the previous module, it remains critical to ensure that LLMs not only align structurally across modalities, but also interpret and reason about these invariant graph structures without introducing biases related to sensitive attributes. To achieve this goal, we utilized the previously aligned sensitive-invariant representations to construct a fairness-aware instruction tuning. Specifically, this paradigm empowers the LLM with robust graph reasoning capabilities while restricting the input graph tokens to those derived solely from sensitive-invariant attributes, thereby ensuring fairness.

To effectively integrate previously grounded invariant embeddings $\hat{\mathbf{E}}^{\text{inv}} \in \mathbb{R}^{N\times d}$ into a pretrained LLM, it is necessary to convert these embeddings into graph tokens compatible with the LLM's input format. In order to efficiently achieve this, we use a lightweight linear projection function: $\mathcal{X}_n^G = L_P(\hat{\mathbf{E}}n^{\text{inv}}), \quad \mathcal{X}_n^G \in \mathbb{R}^{d\text{LLM}}$, where $L_P(\cdot)$ is designed to map invariant graph embeddings into the embedding space of the pretrained LLM encoder $\Phi\text{L}(\cdot)$, thus ensuring minimal computational overhead and compatibility with downstream tasks. The graph tokens are organized sequentially by performing $h$-hop neighbor sampling around a central node, resulting in an ordered token sequence structured as follows: $\mathcal{X}^G(v) = \{\mathbf{x}_{u_1}^G, \mathbf{x}_{u_2}^G, \ldots, \mathbf{x}_{u_m}^G\}, \quad u_i \in \mathcal{M}_h(v)$, where $\mathcal{M}_h(v)$ denotes the nodes sampled from the $h$-hop neighborhood around node $v$. This ordered sequence maintains clear locality information essential for graph reasoning tasks.

To enable the pretrained LLM to effectively interpret and reason about invariant graph tokens without reliance on explicitly labeled data, we further propose a structure-aware self-supervised graph matching task. Specifically, given the invariant graph tokens generated in the previous step, we sample a $h$-hop subgraph $\mathcal{S}(v)$ around each central node $v$ using random neighbor sampling. We then construct a sequential ordering of the invariant graph tokens $\mathcal{X}^G(v)$ according to the sampled subgraph structure. Corresponding textual descriptions, such as titles or node attributes, are collected for each node within this subgraph. To form the self-supervised training task, we randomly shuffle these textual descriptions and instruct the LLM to recover their original order, guided solely by the provided invariant graph tokens. This setup allows the model to inherently capture structural relationships embedded in the tokens, facilitating robust graph reasoning capabilities in the absence of direct supervision.

Given shuffled texts $T' = t_{\pi(1)}, \ldots, t_{\pi(m)}$, where $\pi$ denotes a permutation, we aim to maximize conditional likelihood:

$$\max_{L_P} \sum_{i=1}^{m} \log p\left(t_{\pi^{-1}(i)} \mid \Phi_{\text{L}}(\mathcal{X}^G(v)), t_{\pi^{-1}(1)}, \ldots, t_{\pi^{-1}(i-1)}\right) \tag{12}$$

In particular, during this stage, we freeze the parameters of both the LLM $\Phi_{\text{L}}(\cdot)$ and the invariant graph encoder, training solely the lightweight projector $L_P(\cdot)$. This strategy ensures efficiency and prevents overfitting to specific graph data.

For targeted downstream tasks such as node classification and link prediction, we introduce an instruction fine-tuning stage. Initialized with the pretrained projector $L_P(\cdot)$ from first stage, this stage involves providing the LLM with invariant graph tokens $\mathcal{X}^G(v)$, a concise textual description of the central node, and a clearly defined query specifying the intended task. The LLM is then expected to generate task-specific responses, including predicted labels or structured explanations. We employ the standard negative log-likelihood loss for instruction tuning:

$$\mathcal{L}_{\text{task}} = -\sum_{i=1}^{l} \log p_{L_P}\left(y_i \mid \Phi_{\text{L}}(\mathcal{X}^G(v)), y_{<i}\right) \tag{13}$$

where $y_i$ denotes the tokens of the target sequence and $l$ is the length of the response. For computational efficiency, we freeze both the pretrained LLM encoder and the graph encoder, optimizing only the lightweight projection module. Key hyperparameters, such as neighborhood size $h$, token sequence length, and temperature parameter $\tau$, should be carefully tuned for optimal performance. Overall, by exclusively leveraging sensitive-invariant attribute representations, our proposed instruction tuning equips the LLM to perform fair, robust, and interpretable predictions and classifications

| Method | German | | | | Bail | | | | Credit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Accuracy(↑) | Macro-F1(↑) | SPD(↓) | EOD(↓) | Accuracy(↑) | Macro-F1(↑) | SPD(↓) | EOD(↓) | Accuracy(↑) | Macro-F1(↑) | SPD(↓) | EOD(↓) |
| Vanilla-LLM | 0.579 | 0.680 | 0.205 | 0.191 | 0.514 | 0.523 | 0.149 | 0.141 | 0.643 | 0.722 | 0.191 | 0.167 |
| LLaGA | 0.725 | 0.822 | 0.191 | 0.217 | **0.757** | 0.747 | 0.155 | 0.145 | 0.748 | 0.825 | 0.179 | 0.153 |
| GaLM | 0.715 | **0.828** | 0.187 | 0.208 | 0.731 | 0.728 | 0.140 | 0.149 | 0.751 | 0.814 | 0.184 | 0.138 |
| GraphTranslator | 0.707 | 0.822 | 0.216 | 0.203 | 0.712 | 0.694 | 0.171 | 0.154 | 0.734 | 0.807 | 0.199 | 0.148 |
| GraphLLM | 0.693 | 0.787 | 0.151 | 0.168 | 0.685 | 0.671 | 0.128 | 0.132 | 0.711 | 0.788 | 0.160 | 0.127 |
| GraphICL | **0.733** | 0.825 | 0.177 | 0.164 | 0.742 | **0.751** | 0.148 | 0.143 | **0.763** | **0.840** | 0.156 | 0.131 |
| **FairGEnt** | 0.717 | 0.825 | **0.083** | **0.075** | 0.739 | 0.730 | **0.081** | **0.079** | 0.735 | 0.827 | **0.069** | **0.072** |

Table 1: Comparison of FairGEnt with baseline methods on German, Bail and Credit datasets for node classification task. The best performing result in each row is highlighted in bold, and the second-best result is underlined.

on graph-structured data, effectively mitigating biases associated with sensitive attributes.

# 5 Experiments

## 5.1 Datasets

We evaluate FairGEnt on three widely used real-world datasets, namely **German** [Asuncion and Newman, 2007], **Bail** [Agarwal *et al.*, 2021], and **Credit** [Yeh and Lien, 2009]. The **German** dataset consists of bank clients represented as nodes, where edges capture similarities between individuals and node labels indicate credit risk. The **Bail** dataset is constructed from criminal justice records, where nodes denote defendants and labels correspond to bail outcomes. The **Credit** dataset is derived from credit card records, where nodes represent individuals and labels indicate default risk. For all datasets, node features describe demographic and behavioral attributes, while gender, race, and age are used as the sensitive attributes, respectively. The detailed statistics of these datasets are reported in Table 2.

## 5.2 Baselines

We compare the proposed method with several state-of-the-art baselines, categorized into two groups: i) Vanilla Language Model, which aim to improve LLM reasoning by injecting graph information through different integration strategies. Specifically, Vanilla-LLM, a standard LLM without explicit integration of graph data and fairness considerations; ii) Graph-enhanced Language Models: GaLM [Xie *et al.*, 2023] utilizes graph-attention mechanisms to jointly learn graph and text embeddings; LLaGA [Chen *et al.*, 2024] reorganizes graph nodes into structured token sequences, enabling pretrained LLMs to directly reason over graph data; GraphTranslator [Zhang *et al.*, 2024a] transforms structured graph data into textual representations through a translation-based framework, thereby bridging graph inputs and language modeling; GraphLLM [Chai *et al.*, 2025] employs vector quantization to tokenize graph structures, enabling LLMs to directly leverage graph information; and GraphICL [Sun *et al.*, 2025] integrates graph-structured information into language models by leveraging in-context learning, enabling effective structural reasoning without explicit fine-tuning or parameter updates.

| Dataset | German | Bail | Credit |
|---|---|---|---|
| # Nodes | 1,000 | 18,876 | 30,000 |
| # Edges | 22,242 | 311,870 | 137,377 |
| # Features | 27 | 18 | 13 |
| # Average Degree | 44.48 | 34 | 10 |
| Sensitive Attribute | Gender | Race | Age |

Table 2: Summary of the datasets used in the experiments.

## 5.3 Evaluation Metrics

We evaluate our framework from two perspectives: predictive performance and fairness. For predictive performance, we adopt Accuracy and Macro-F1 [Sokolova and Lapalme, 2009], where higher values indicate better classification performance. For fairness evaluation, we employ two widely used group fairness metrics: Statistical Parity Difference (SPD) [Dwork *et al.*, 2012] and Equal Opportunity Difference (EOD) [Hardt *et al.*, 2016]. These metrics quantify disparities in model predictions across different sensitive groups, with values closer to zero indicating improved fairness.

## 5.4 Experiment Results

**Comparison Study.** We compare FairGEnt against six baseline models on three datasets. The averaged results are summarized in Table 1. The experimental outcomes demonstrate that FairGEnt consistently achieves superior performance compared to baseline methods across most evaluation metrics. Specifically, FairGEnt significantly outperforms all baselines in fairness metrics, indicating its strong capability in mitigating bias propagation within graph-enhanced LLMs. The notable fairness improvements can be primarily attributed to three key factors: (i) explicitly disentangling sensitive attributes from task-relevant information in graph and textual embeddings, thereby reducing biased correlations during representation learning; (ii) employing a fairness-guided alignment module that integrates disentangled embeddings with LLM representations, effectively minimizing residual

bias leakage; and (iii) refining the fairness-aware representation through a graph-enhanced instruction tuning process, which promotes unbiased downstream reasoning. Moreover, FairGEnt maintains competitive predictive performance comparable to or surpassing the baselines, highlighting the framework's ability to simultaneously ensure fairness and retain utility performance. Overall, these results confirm the effectiveness of FairGEnt in delivering fair and robust predictions within graph-structured language modeling tasks.

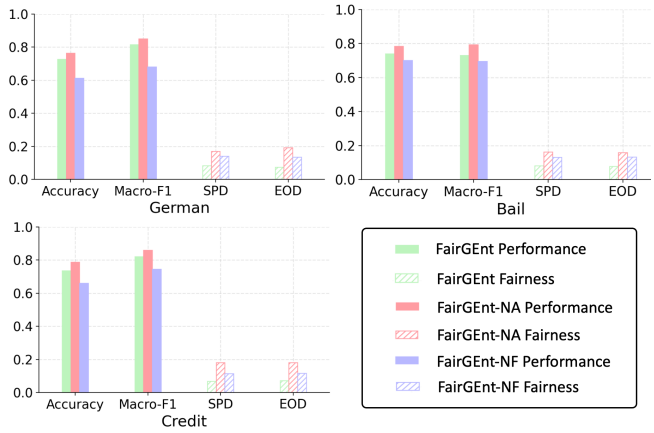


Figure 1: Ablation study results for FairGEnt, FairGEnt-NA and FairGEnt-NF.

**Ablation Study.** We conducted ablation studies to investigate the contributions of individual modules within FairGEnt. Specifically, we constructed two variants: FairGEnt-NA and FairGEnt-NF. FairGEnt-NA integrates graph-enhanced embeddings with LLM representations without applying explicit disentanglement of sensitive attributes, thereby prioritizing task-relevant information and predictive performance but omitting fairness constraints. In contrast, FairGEnt-NF incorporates sensitive attribute disentanglement but excludes the fairness-guided alignment module, thus lacking explicit fairness constraints during representation integration. Results are summarized in Figure 1. Compared to FairGEnt, FairGEnt-NA achieves higher predictive performance but significantly compromised fairness, as sensitive attributes remain entangled with task-relevant information, leading to substantial bias propagation. In contrary, FairGEnt-NF exhibits decreased predictive performance due to potential loss of task-relevant information during disentanglement; however, fairness metrics notably improve relative to FairGEnt-NA, highlighting the effectiveness of sensitive attribute disentanglement. Finally, our proposed FairGEnt demonstrates optimal fairness; at the same time, its prediction performance is better than FairGEnt-NF, validating that combining sensitive attribute disentanglement and fairness-guided alignment effectively balances performance and fairness.

**Hyperparameter Sensitivity Analysis.** We further conduct a hyperparameter sensitivity analysis to examine how FairGEnt balances fairness preservation and predictive utility under different strengths of disentanglement and cross-modal alignment. In the overall pre-training objective, $\alpha$ controls the contribution of the graph and text disentanglement losses,

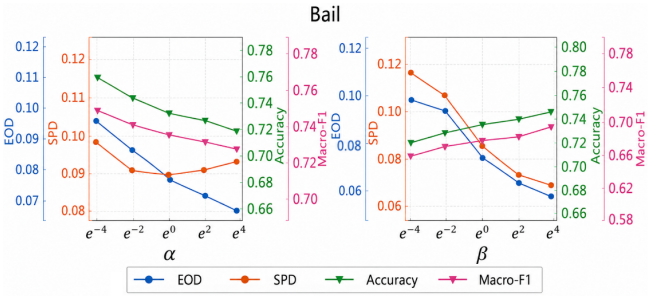


Figure 2: Hyperparameters sensitivity study results on the Bail dataset.

while $\beta$ regulates the invariant graph–text alignment objective. Therefore, these two parameters directly determine how strongly FairGEnt suppresses sensitive information and how effectively it transfers sensitive-invariant knowledge across modalities. We vary $\alpha$ and $\beta$ over the exponential range $\{e^{-4}, e^{-3}, \ldots, e^{3}, e^{4}\}$ on the Citeseer dataset, and report the results in Figure 2. The results show that parameters play a meaningful role in shaping the fairness–utility trade-off. As $\alpha$ increases, the model places greater emphasis on separating sensitive-related factors from invariant representations, leading to lower SPD and EOD. However, overly large values may also remove part of the task-relevant signal that is statistically entangled with sensitive attributes, resulting in a mild decline in Accuracy and Macro-F1. Similarly, increasing $\beta$ improves fairness by encouraging stronger alignment between invariant graph and textual subspaces, thereby reducing bias propagation during cross-modal integration. Nevertheless, excessive alignment strength can constrain representation flexibility and slightly affect predictive performance. Overall, the best results are achieved when $\alpha$ and $\beta$ are set to moderate values, suggesting that FairGEnt benefits most from a balanced optimization strategy, sufficiently strong disentanglement to suppress sensitive leakage, together with stable invariant alignment to preserve graph-text semantic consistency.

## 6 Conclusion

Given the inherent limitations of existing LLMs in effectively and fairly processing graph-structured data, this paper presents an initial exploration into fairness-aware integration of graph knowledge and language modeling. Specifically, we propose *FairGEnt*, a disentangled graph-enhanced large language model designed explicitly to mitigate biases from structural and textual sources. FairGEnt disentangles sensitive-related and sensitive-invariant factors, thereby preventing protected attributes from adversely influencing downstream predictions. Additionally, it incorporates fairness-aware alignment and instruction-tuning strategies to enable unbiased integration and enhance graph-structure comprehension. Extensive empirical evaluations demonstrate that FairGEnt consistently outperforms baseline methods in fairness and predictive accuracy, confirming its effectiveness as a practical and generalizable approach to fair graph-enhanced language modeling.

## References

[Agarwal *et al.*, 2021] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, pages 2114–2124. PMLR, 2021.

[Asuncion and Newman, 2007] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[Azher *et al.*, 2024] Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–12, 2024.

[Bonner *et al.*, 2022] Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L Hamilton. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics*, 23(6):bbac404, 2022.

[Brandes *et al.*, 2013] Ulrik Brandes, Markus Eiglsperger, Jürgen Lerner, and Christian Pich. Graph markup language (graphml). 2013.

[Campbell *et al.*, 2013] William M Campbell, Charlie K Dagli, and Clifford J Weinstein. Social network analysis with content and graphs. *Lincoln Laboratory Journal*, 20(1):61–81, 2013.

[Chai *et al.*, 2025] Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiang Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. Graphllm: Boosting graph reasoning ability of large language model. *IEEE Transactions on Big Data*, 2025.

[Chen *et al.*, 2024] Runjin Chen, Tong Zhao, Ajay Jaiswal, Neil Shah, and Zhangyang Wang. Llaga: Large language and graph assistant. *arXiv preprint arXiv:2402.08170*, 2024.

[Chinta *et al.*, 2024] Sribala Vidyadhari Chinta, Zichong Wang, Zhipeng Yin, Nhat Hoang, Matthew Gonzalez, T Le Quy, and Wenbin Zhang. Fairaied: Navigating fairness, bias, and ethics in educational ai applications. *arXiv preprint arXiv:2407.18745*, 2024.

[Chu *et al.*, 2024] Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1):34–48, 2024.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Gong *et al.*, 2024] Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18362–18372, 2024.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[Koroteev, 2021] Mikhail V Koroteev. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021.

[Li *et al.*, 2025] Mengran Li, Pengyu Zhang, Wenbin Xing, et al. A survey of large language models for data challenges in graphs. *arXiv preprint arXiv:2505.18475*, 2025. Survey on LLM-graph integration challenges.

[Lopes *et al.*, 2020] Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. Lite training strategies for Portuguese-English and English-Portuguese translation. In *Proceedings of the Fifth Conference on Machine Translation*, Online, November 2020. Association for Computational Linguistics.

[Mienye, 2025] ID Mienye. Large language models: an overview of foundational architectures, training methodologies, and applications. *Springer Journal of AI Research*, 2025. Review of LLM foundations and applications.

[Osman and Barukub, 2020] Ahmed Hamza Osman and Omar Mohammed Barukub. Graph-based text representation and matching: A review of the state of the art and future challenges. *IEEE Access*, 8:87562–87583, 2020.

[Pourhabibi *et al.*, 2020] Tahereh Pourhabibi, Kok-Leong Ong, Booi H Kam, and Yee Ling Boo. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133:113303, 2020.

[Radaideh *et al.*, 2025] Mohammed I Radaideh, O Hwang Kwon, and Majdi I Radaideh. Fairness and social bias quantification in large language models for sentiment analysis. *Knowledge-Based Systems*, page 113569, 2025.

[Sokolova and Lapalme, 2009] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[Sun *et al.*, 2025] Yuanfu Sun, Zhengnan Ma, Yi Fang, Jing Ma, and Qiaoyu Tan. Graphicl: Unlocking graph learning potential in llms through structured prompt design. *arXiv preprint arXiv:2501.15755*, 2025.

[Tang *et al.*, 2024] Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. Higpt: Heterogeneous graph language model. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2842–2853, 2024.

[Wang *et al.*, 2025a] Zichong Wang, Avash Palikhe, Zhipeng Yin, and Wenbin Zhang. Fairness in language models: A tutorial. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6849–6852, 2025.

[Wang *et al.*, 2025b] Zichong Wang, Zhipeng Yin, Zhen Liu, Roland HC Yap, Xiaocai Zhang, Shu Hu, and Wenbin

Zhang. Redefining fairness: A multi-dimensional perspective and integrated evaluation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 336–353. Springer, 2025.

[Wang *et al.*, 2025c] Zichong Wang, Zhipeng Yin, Liping Yang, Jun Zhuang, Rui Yu, Qingzhao Kong, and Wenbin Zhang. Fairness-aware graph representation learning with limited demographic information. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 354–371. Springer, 2025.

[Wang *et al.*, 2025d] Zichong Wang, Zhipeng Yin, Roland HC Yap, and Wenbin Zhang. Ai fairness beyond complete demographics: Current achievements and future directions. *arXiv preprint arXiv:2511.13525*, 2025.

[Wang *et al.*, 2026a] Zichong Wang, Zhipeng Yin, Zhong Chen, Jack Yang, Jun Liu, and Wenbin Zhang. Learning counterfactual fairness from authentic generation. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence*, 2026.

[Wang *et al.*, 2026b] Zichong Wang, Zhipeng Yin, Mo Sha, Xiaofeng Gao, Xiaoli Li, and Wenbin Zhang. Towards fair graph learning without demographic supervision. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence*, 2026.

[Wang *et al.*, 2026c] Zichong Wang, Zhipeng Yin, and Wenbin Zhang. A unified framework for fair graph generation: Theoretical guarantees and empirical advances. *Advances in Neural Information Processing Systems*, 38, 2026.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Xie *et al.*, 2023] Han Xie, Da Zheng, Jun Ma, Houyu Zhang, Vassilis N Ioannidis, Xiang Song, Qing Ping, Sheng Wang, Carl Yang, Yi Xu, et al. Graph-aware language model pre-training on a large graph corpus can help multiple graph applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5270–5281, 2023.

[Yang *et al.*, 2025] Xihong Yang, Heming Jing, Zixing Zhang, Jindong Wang, Huakang Niu, Shuaiqiang Wang, Yu Lu, Junfeng Wang, Dawei Yin, Xinwang Liu, et al. Darec: A disentangled alignment framework for large language model and recommender system. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 904–917. IEEE, 2025.

[Yeh and Lien, 2009] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

[Yi *et al.*, 2024] Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*, 2024.

[Yin *et al.*, 2025] Zhipeng Yin, Zichong Wang, Avash Palikhe, Zhen Liu, Jun Liu, and Wenbin Zhang. Amcr: A framework for assessing and mitigating copyright risks in generative models. *28th European Conference on Artificial Intelligence*, 2025.

[Yin *et al.*, 2026] Zhipeng Yin, Zichong Wang, Ruijun Chen, Xin Ning, Xingyu Zhang, and Wenbin Zhang. Toward LoRA copyright protection with an authorized dual-watermarking framework. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence*, 2026.

[Zhang and Ntoutsi, 2019] Wenbin Zhang and Eirini Ntoutsi. Faht: An adaptive fairness-aware decision tree classifier. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1480–1486. International Joint Conferences on Artificial Intelligence Organization, 2019.

[Zhang and Weiss, 2022] Wenbin Zhang and Jeremy C Weiss. Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, 2022.

[Zhang *et al.*, 2023] Wenbin Zhang, Tina Hernandez-Boussard, and Jeremy Weiss. Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, 2023.

[Zhang *et al.*, 2024a] Mengmei Zhang, Mingwei Sun, Peng Wang, Shen Fan, Yanhu Mo, Xiaoxiao Xu, Hong Liu, Cheng Yang, and Chuan Shi. Graphtranslator: Aligning graph model to large language model for open-ended tasks. In *Proceedings of the ACM Web Conference 2024*, pages 1003–1014, 2024.

[Zhang *et al.*, 2024b] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, 2024.

[Zhang *et al.*, 2025a] Wenbin Zhang, Shuigeng Zhou, Toby Walsh, and Jeremy C Weiss. Fairness amidst non-iid graph data: A literature review. *AI Magazine*, 2025.

[Zhang *et al.*, 2025b] Yazhou Zhang, Mengyao Wang, Qiuchi Li, Prayag Tiwari, and Jing Qin. Pushing the limit of llm capacity for text classification. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1524–1528, 2025.

[Zhang, 2024] Wenbin Zhang. Fairness with censorship: Bridging the gap between fairness research and real-world deployment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Zmigrod *et al.*, 2019] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics.