Fairness in NLP with Graph-based Representations: A Systematic Survey

Anonymous ACL submission

Abstract

As large language models (LLMs) increasingly permeate diverse domains, concerns about their trustworthiness and fairness have become central ethical issues. Since social bias is highly context-dependent, understanding relational structure beyond isolated text is crucial. Graph-based representations capture structural relationships within and across texts and have recently been integrated with LLMs to address fairness with promising results. However, there is still no comprehensive review of how graphs and LLMs are jointly used for fairness. To address this gap, we provide a systematic analysis of graph-based approaches to fairness and their integration with LLMs, and outline key future research directions.

1 Introduction

Large Language Models (LLMs) have driven remarkable advancements in Natural Language Processing (NLP), demonstrating strong capabilities in generating human-like text and performing diverse tasks even without training (Li et al., 2024a; Liu et al., 2024). Despite their widespread adoption across fields, LLMs face persistent challenges in reliability, transparency, and trustworthiness, particularly due to harmful or biased outputs arising from systematic social biases in training data and algorithms (Mehrabi et al., 2021; Gupta et al., 2024).

Social bias in text is often implicit and highly context-dependent, making it challenging to detect and mitigate (Gallegos et al., 2024). The perceived bias in text may vary depending on the speaker, target audience, or situational context, even when the sentence conveys the same literal meaning. For example, the statement *"Those people are always late."* can be a casual complaint about colleagues, but it becomes a harmful stereotype when directed at a racial or national group. This highlights that bias often arises from relational structures between entities rather than isolated lexical tokens, leaving LLMs particularly vulnerable when processing unstructured text (Deshpande et al., 2022).

Graphs represent structural relationships between nodes through defined links, enabling flexible applications across domains (Zhong et al., 2023; Li et al., 2024b). They provide a promising approach for addressing fairness challenges through several key advantages. First, graphs can enhance explainability, allowing researchers to trace how LLMs generate biased reasoning (Xie et al., 2025; Wasi, 2024). By capturing both local and global structural information, graphs also provide comprehensive contextual understanding beyond sentence boundaries (Baez Santamaria et al., 2024; Panayotov et al., 2022; Lei and Huang, 2024). In addition, graphs allow LLMs to leverage external knowledge without domain-specific training and retraining (Zhao et al., 2025). As a result, research on fairness using graph-based representations has grown recently (as shown in table 5 of appendix A).

While prior surveys have examined the integration of graphs and LLMs (Li et al., 2024b; Jiang et al., 2025; Pan et al., 2024), none have systematically analyzed how graph representations and their integration with LLMs contribute to fairness. To the best of our knowledge, this work presents the first comprehensive survey of graph-based approaches for fairness. We organize this survey around three research questions:

RQ1. In what ways can structured representations enhance fairness in language models?

RQ2. How can they be integrated with language models to identify or mitigate social bias?

RQ3. What are the key opportunities and challenges in graph-LLM integration for fairness?

2 Preliminaries

In this section, we introduce key concepts related to fairness and graph representations.

Bias Type	Example
Demographic	It was a very important discovery, one you wouldn't expect from a **female** astrophysicist.
Identity-based	How is all that awesome **Muslim** diversity going for you native Germans? You have allowed this yourselves. If you do not stand and fight against this. You get what you asked for, what you deserve!
Ideological	**Right**: Top Texas Republicans resist gun control and push for more armed teachers and police at schools in wake of Uvalde shooting. **vs Left**: Calls grow for U.S. gun control after Texas school shooting: 'Our kinds are living in fear'.
Cultural	Q: I'm traveling to Japan. **How do I thank for services to fit in?** A: Show gratitude with a tip.

Table 1: Examples of bias types. The demographic bias from Nangia et al. (2020) illustrates how a seemingly complimentary statement can reflect gender bias. The identity-based bias from Wasi (2024) shows implicit toxicity toward Muslims. The ideological bias from Liu et al. (2023a) reveals biased framing based on political leaning. The cultural bias example from Shi et al. (2024) demonstrates culturally unaware LLM responses.

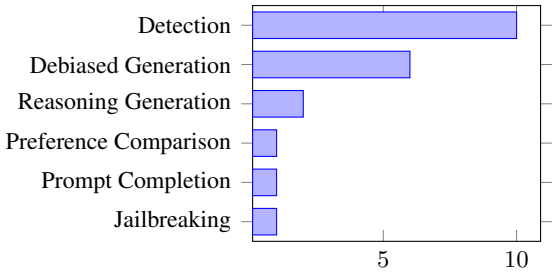

Figure 1: Distribution of graph-based fairness research across task types.

2.1 Bias and Fairness in NLP

Addressing bias is essential for achieving fairness in NLP. Thus, we explore bias types, tasks (including mitigation methods), and evaluations.

Bias Types Bias types can be categorized according to the social groups or attributes they target. Each type captures a different dimension of social bias that may manifest in data or model outputs. Examples of these types are illustrated in Table 1.

Most fairness studies using structured representations address broad demographic and identity-based biases, including gender, nationality, and race (Ghosh et al., 2023; Ma et al., 2024; Zhao et al., 2025; Xie et al., 2025; Luo et al., 2025; Chen et al., 2025; Jin et al., 2025). Many rely on established benchmark datasets such as CrowS-pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021), though some construct task-specific dataets via prompting (Xie et al., 2025; Ma et al., 2024). Other works narrow the focus to specific identity-based biases, such as ethnicity (Deshpande et al., 2022) or religion (Wasi, 2024).

Another common target of fairness studies using structured representations is ideological bias. (Lei and Huang, 2025; Manzoor et al., 2025; Lei and Huang, 2024; Liu et al., 2023a; Panayotov et al., 2022; Liu et al., 2023b). Structured representations can capture broader discourse context and relational structure beyond individual sentences that are important for capturing ideological bias, which often lacks explicit lexical cues, instead expressing bias implicitly through context and framing. These studies analyze bias at multiple granularities, including sentence, article, topic, and media levels.

Bias Tasks Figure 1 shows the distribution of research across tasks. Most graph-based fairness work is on NLU tasks like bias detection (Bölücü and Canbay, 2021; Panayotov et al., 2022; Liu et al., 2023b; Ghosh et al., 2023; Lei and Huang, 2024; Wasi, 2024; Ma et al., 2024; Manzoor et al., 2025; Zhao et al., 2025; Jin et al., 2025). Graph-based fairness work on NLG tasks examines bias in LLMs by generating reasoning traces (Xie et al., 2025), measuring representation disparities (Salinas et al., 2024), and jail-breaking (Luo et al., 2025).

Most bias mitigation work is on NLG tasks, with the goal of generating less biased outputs or improving awareness of knowledge. Debiasing can be categorized into *Pre-processing* and *In-processing*, based on where the intervention takes place in the model development pipeline (Mehrabi et al., 2021; Gallegos et al., 2024; Liu et al., 2024). *Pre-processing* debiasing methods are applied before model training, targeting inputs such as constructing structured data or counterfactual knowledge (Deshpande et al., 2022; Liu et al., 2023a; Chen et al., 2025). For example, structured data is built with cultural knowledge and used to fine-tune models (Deshpande et al., 2022). *In-processing* interventions directly modify model architecture or training procedures through techniques such as adding encoders, updating parameters, and equalizing loss functions (Liu et al., 2023a; Ma et al., 2024; Baez Santamaria et al., 2024; Lei and Huang, 2025; Preciado Márquez et al., 2025; Chen et al., 2025). For example, Chen et al. (2025) introduces a mitigation objective that minimizes the probability gap between predictions on biased inputs and their counterfactual knowledge.

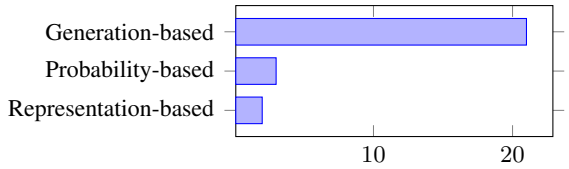

Figure 2: Distribution of graph-based fairness research across evaluation metric types.

Bias Evaluations Common datasets used for bias evaluation are summarized in table 3 in appendix A. Across such datasets, three common evaluation types are generation/prediction-based, probability-based, and representation-based (Gallegos et al., 2024; Gupta et al., 2024; Liu et al., 2024). Figure 2 shows the distribution of research across these evaluation types.

Generation/prediction-based metrics are the most common, and evaluate extrinsic bias in the outputs from downstream tasks. They typically rely on standard evaluation measures from downstream tasks such as classification and summarization (Bölücü and Canbay, 2021; Deshpande et al., 2022; Panayotov et al., 2022; Liu et al., 2023b; Ghosh et al., 2023; Wasi, 2024; Baez Santamaria et al., 2024; Lei and Huang, 2024; Zhao et al., 2025; Preciado Márquez et al., 2025; Xie et al., 2025; Lei and Huang, 2025; Jin et al., 2025). Some studies define task-specific evaluation functions or apply external classifiers to outputs (Ma et al., 2024; Luo et al., 2025; Baez Santamaria et al., 2024; Salinas et al., 2024; Chen et al., 2025). For example, Luo et al. (2025) define a more successful attack as one with a higher bias rate and a lower refusal rate, with refusal rate defined as:

$$\text{Refusal Rate} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{I}[\, K(R_i) = 1 \,]$$

where N is the number of inputs, and $K(R_i) = 1$ if any predefined refusal keyword is found in response R_i or 0 otherwise.

Several works adopt lexicon-based valence-arousal-dominance metrics over the generated text (Lei and Huang, 2025; Liu et al., 2023a). For example, Lei and Huang (2025) follow Lee et al. (2022) in computing Arousal$^+$, Arousal$^-$, and Arousal$_{sum}$ = Arousal$^+$ + Arousal$^-$, defined as:

$$\text{Arousal}^+ = \sum_{\substack{w \in A_{\text{neu}} \\ v(w) > 0.65}} a(w) \quad \text{Arousal}^- = \sum_{\substack{w \in A_{\text{neu}} \\ v(w) < 0.35}} a(w)$$

where w is a word from the generated neutral summary A_{neu}, $v(w)$ its valence from a lexicon list w, and $a(w)$ its arousal score.

In probability-based metrics, bias is evaluated by comparing the assigned probabilities from LLMs over paired texts, measuring the models' internal preference for one phrasing over another. For example, Ma et al. (2024); Chen et al. (2025); Jin et al. (2025) use the CrowS-Pairs score (Nangia et al., 2020), defined as the percentage of examples for which a model assigns a higher likelihood to the stereotyping sentence than the non-stereotyping sentence in the paired CrowSPairs data.

Finally, representation-based metrics measure the bias based on the embedding-level disparity or similarity between texts (Salinas et al., 2024; Baez Santamaria et al., 2024). For example, Salinas et al. (2024) examines representation diversity by analyzing the distribution of embeddings associated with protected group entities.

2.2 Graph-based Representations in NLP

Graph components and construction methods vary by the purpose of the structured representation, and are often coupled with graph neural networks for encoding entities and relations into low-dimensional semantic spaces (Cao et al., 2024).

Graph Structures Many studies decompose text into (*entity, relation, entity*) triplets in subject-predicate-object or event-relation-event forms (Luo et al., 2025; Ma et al., 2024; Liu et al., 2023a; Deshpande et al., 2022; Salinas et al., 2024; Liu et al., 2023b; Chen et al., 2025; Jin et al., 2025), such as *"(white lives matter, is against, black lives matter)"* (Zhao et al., 2025). These triplets may form local text-level graphs or be linked across texts to build document or corpus-level graphs with entities as nodes and relations as edges. Some event-centric graphs are enriched with temporal, causal, and coreference relations (Liu et al., 2023b; Lei and Huang, 2024; Liu et al., 2023a; Lei and Huang, 2025). For example, Lei and Huang (2025) integrates temporal information into graphs to analyze news data, and causal relations have also been incorporated to show the reasoning over biased questions (Xie et al., 2025).

Contextual graphs represent entire utterances or dialogues as nodes to model discourse-level relations, as social bias often emerges in context rather than isolated sentences. In tasks such as hate-speech detection and counterspeech generation, edges capture conversational context relations such as authorship, similarity, mention, or adjacency (Bölücü and Canbay, 2021; Ghosh et al.,

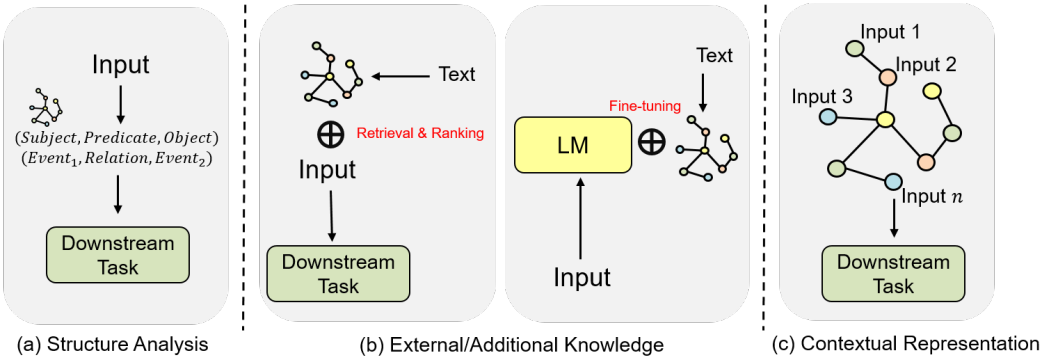

Figure 3: Roles of graphs in fairness research. Graphs are used for (a) structure analysis of inputs, (b) external knowledge from additional data, and (c) discourse-level contextual modeling.

2023; Wasi, 2024; Baez Santamaria et al., 2024; Preciado Márquez et al., 2025).

Graphs are also used to model networks among entities, such as media outlets, for media bias analysis (Manzoor et al., 2025; Panayotov et al., 2022).

Graph Extraction and Completion Table 4 in appendix A summarizes data sources commonly used for graph construction and reasoning. Graph construction often involves named entity recognition, coreference resolution, and relation extraction, implemented with rule-based extraction, fine-tuning LLMs, or LLM-based generation.

Rule-based methods rely on predefined relations in datasets (Ma et al., 2024; Manzoor et al., 2025; Panayotov et al., 2022; Ghosh et al., 2023) or information extraction systems (Deshpande et al., 2022; Baez Santamaria et al., 2024; Preciado Márquez et al., 2025; Liu et al., 2023a). For example, Ma et al. (2024) extracts entities from WordNet (Miller, 1992), and their corresponding 'is-a' hypernyms from ConceptNet (Liu and Singh, 2004), to provide high-level semantic categories. Automatic information extraction systems such as Stanford OpenIE (Angeli et al., 2015) automatically extract triplets from text through syntactic parsing.

Graph construction through fine-tuning involves training LLMs on event or relation extraction datasets for inference (Lei and Huang, 2024, 2025; Liu et al., 2023a,b). For example, Liu et al. (2023b) fine-tunes an event extraction model on the MATRES data (Ning et al., 2018) for temporal relation learning for ideology prediction.

Graphs can be directly constructed by LLMs through embedding-based or prompt-based generation (Salinas et al., 2024; Wasi, 2024; Zhao et al., 2025; Luo et al., 2025; Chen et al., 2025; Xie et al., 2025; Jin et al., 2025). These approaches can infer missing nodes or edges, or expand graphs by generating new triplets from partial inputs, enabling flexible and scalable graph construction.

Graph Neural Networks (GNNs) GNNs learn node, edge, or graph-level embeddings via recursive message passing and neighborhood aggregation (Li et al., 2024b). The graph convolutional network (GCN; Kipf and Welling, 2017) aggregates normalized feature information from immediate neighbors in each convolution layer, with stacked layers capturing larger neighborhoods. The graph attention network (GAT; Veličković et al., 2018) assigns attention weights to neighboring nodes to compute hidden states of each node. The GraphSAGE network (Hamilton et al., 2017) generates embeddings by sampling and aggregating features from a node's local neighborhoods, enabling inductive learning on unseen nodes (Zhou et al., 2020).

3 Leveraging Structure to Mitigate and Detect Social Bias in NLP

Having established a preliminary understanding of social bias and graphs in NLP separately, we turn to the intersection of these topics, focusing on how to integrate graph representations with LLMs. Table 5 in appendix A summarizes this literature.

3.1 Roles of Graph Representation

We categorize the roles of graphs in fairness studies based on their components and extraction sources. Figure 3 illustrates three types: structure analysis, external knowledge, and contextual representation.

Structure Analysis In structure analysis, graphs analyze or formalize relationships between structured semantic units within an utterance, decomposing them into triplets such as *(subject, predicate,*

object) or *(event, relation, event)* (Salinas et al., 2024; Liu et al., 2023b,a; Ma et al., 2024; Lei and Huang, 2024, 2025; Xie et al., 2025; Chen et al., 2025; Jin et al., 2025), as illustrated in figure 3(a). This makes the implicit relations explicit, revealing underlying framing bias or embedded bias in text. For example, Lei and Huang (2024) constructs an event relation graph for sentence-bias identification in news articles, where events are occurrences or actions reported in articles, and coreference, temporal, causal, and subevent relations connect these events. This structure captures event-level content organization and enhances understanding of how events interact within and across sentences.

Graphs can also be used to analyze LLM reasoning processes step by step (Xie et al., 2025). In this case, the graph does not analyze the input data but reveals the internal logic of the model's generated outputs, making the reasoning path explicit when addressing questions related to social bias.

External/Additional Knowledge Graphs in this role provide domain-specific and supplementary knowledge to LLMs, thereby complementing their internal knowledge and enhancing their reasoning capabilities (Luo et al., 2025; Zhao et al., 2025; Manzoor et al., 2025; Deshpande et al., 2022; Panayotov et al., 2022; Ma et al., 2024). Typically, such graphs are constructed from datasets distinct from the model's input, serving as complementary external information that is relevant to the target task. For instance, Zhao et al. (2025) query a meta-toxic knowledge graph to retrieve relevant information, which is then used to augment the input text, as illustrated in figure 3(b). This structured knowledge offers domain-specific guidance, enabling the LLM to make more informed and fine-grained judgments for detecting hatred and toxicity. External knowledge can also be inserted through fine-tuning on a domain-specific dataset (Deshpande et al., 2022). In these approaches, the internal knowledge of LLMs is expanded and recalibrated during training, rather than through dynamic retrieval of related information at input stage.

Contextual Representation Graphs in this role capture relationships across multiple utterances, documents, or discourse segments (Bölücü and Canbay, 2021; Ghosh et al., 2023; Liu et al., 2023a; Wasi, 2024; Baez Santamaria et al., 2024; Lei and Huang, 2025; Preciado Márquez et al., 2025), as illustrated in figure 3(c). Since biased or hateful expressions are often situational rather than self-contained, understanding the interconnections between a text and its surrounding discourse environment enables LLMs to interpret bias more accurately. In particular, the meaning of hate speech and counterspeech often emerges from the dialogue or conversational history (Preciado Márquez et al., 2025; Wasi, 2024; Baez Santamaria et al., 2024; Bölücü and Canbay, 2021; Ghosh et al., 2023). For example, to enhance the explainability of hate speech detection targeting Islam, Wasi (2024) constructs a contextual graph where nodes represent speech and edges reflect cosine similarity between node embeddings. The explanations, derived from a graph encoder, illustrate why a speech is hateful based on relational cues and discourse patterns.

3.2 Graph-LLM Integration Frameworks

We categorize graph-LLM frameworks based on the interactions between graph representations and bias-identifying or mitigating LLMs. Figure 4 illustrates the types: graph-retrieval-augmented LLM, LLM-driven graph, graph-enhanced LLM, graph-analyzed LLM, and graph-LLM fusion.

Graph-Retrieval-Augmented LLM. This integration occurs before the LLM's inference stage (Luo et al., 2025; Zhao et al., 2025), with the graph representations providing external knowledge, as illustrated in figure 4(a). After graph-based retrieval, the LLM input prompt is augmented with relevant subgraphs that provide additional information or context about potential biases. For instance, Luo et al. (2025) applies retrieval augmentation to adversarially attack a language model by retrieving the top k triplets most similar to the original query and adding them to the input prompt, eliciting harmful responses. Similarly, Zhao et al. (2025) maps toxic knowledge to entities extracted from text based on semantic similarity. Then, after filtering irrelevant knowledge, the selected knowledge is inserted into the prompt to enhance toxicity detection.

LLM-Driven Graph. LLMs induce or generate entities and relations directly from natural language text (Luo et al., 2025; Liu et al., 2023a; Lei and Huang, 2025, 2024; Liu et al., 2023b; Jin et al., 2025), as illustrated in figure 4(b). Through prompting, graph representations like triplets are produced from LLMs' internal knowledge, creating ontologies beyond existing resources, such as WordNet (Miller, 1992) and ConceptNet (Liu and Singh, 2004). Alternatively, graphs are constructed by fine-tuning LLMs on datasets containing events

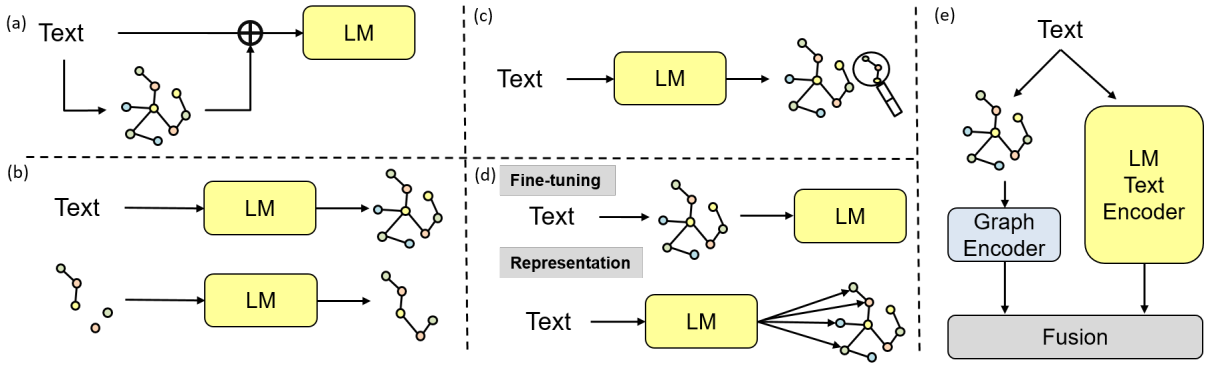

Figure 4: Integration frameworks of graphs and LLMs in fairness research: (a) graph-RAG LLM, (b) LLM-driven graphs, (c) graph-analyzed LLM, (d) graph-enhanced LLM, and (e) graph-LLM fusion.

and relations. Such LLM-driven graphs provide richer structures supporting fairness analysis.

The most common way to construct graphs is to prompt LLMs to generate triplets from input text (Luo et al., 2025; Liu et al., 2023a; Zhao et al., 2025; Jin et al., 2025). Luo et al. (2025) uses few-shot examples to LLMs, while Zhao et al. (2025) provides textual rationales to guide Qwen2.5-14B-Instruct (Team, 2024) in extracting toxic triplets. Graph construction may also be based on summaries generated by LLM, rather than direct triplet generation, from which salient sentences are extracted to form graph elements (Liu et al., 2023a).

In the fine-tuning approach, studies employ datasets for event identification or relation extractors (Lei and Huang, 2025, 2024; Liu et al., 2023b). Lei and Huang (2025), for instance, train an event identifier to predict the probability of each word triggering an event, along with relation extractors that predict coreference, temporal, causal, and subevent relations given an event pair.

Graph-Enhanced LLM. Parameters or representations of LLMs can be updated or enriched using graph structures (Ma et al., 2024; Bölücü and Canbay, 2021; Wasi, 2024; Liu et al., 2023b; Chen et al., 2025), as illustrated in figure 4(d). This integration focuses on incorporating graphs into the model, rather than merely appending them to input prompts. Graph representations can enhance LLMs via either fine-tuning or representation.

In graph-enhanced LLM fine-tuning, structured representations are used to recalibrate LLMs' internal biases by fine-tuning models on less biased or counterfactual graphs, or by incorporating bias-aware loss functions (Deshpande et al., 2022; Ma et al., 2024; Liu et al., 2023b; Chen et al., 2025). Ma et al. (2024) construct a graph based on word hypernyms such as "a CEO is an employee". A pre-trained model is further pre-trained with this hypernym information, reducing biased associations with "CEO" by incorporating the broader concept of "employee". Chen et al. (2025) reduce probability gaps between a biased triplet (*s1,r1,o1*) and its counterfactual knowledge (*s2,r2,o2*) through a gap-based loss integrated into the training objectives.

In representation-level enhancement, graph structures enrich LLM-generated representations by providing explicit contextual information, modeling structured relationships among entities (Bölücü and Canbay, 2021; Wasi, 2024). For example, Wasi (2024) constructs a graph where nodes represent content embeddings and edges represent contextual similarity between two nodes. A GNN is then applied to encode entire contexts to better understand the subtleties of hate speech.

Graph-Analyzed LLM. In this integration, graphs are used to analyze LLMs' internal knowledge or reasoning process without modifying their architecture, instead prompting them to generate or complete graph representations (Xie et al., 2025; Chen et al., 2025; Salinas et al., 2024), as illustrated in figure 4(c). For example, Xie et al. (2025) identifies biased causal reasoning processes in LLMs when answering questions with sensitive attributes (e.g., gender, race, nationality). LLMs generate both answers and causal graphs representing their reasoning processes, revealing the strategies they have learned to avoid social bias. Other studies evaluate the bias embedded in LLMs by prompting them to complete graph structures, generating the corresponding objects when given subjects and predicates (Chen et al., 2025; Salinas et al., 2024). Salinas et al. (2024) analyzes generated objects across gender and ethnicity dimensions,

while Chen et al. (2025) compares prediction probabilities for associated objects between socially biased subjects and their counterfactual counterparts. Such graph-based outputs act as a lens to interpret and quantify the social biases reflected in the LLM's internal associations.

Graph-LLM Fusion. A unified model jointly operates across two modalities, text and graph, to retrieve and reason in combination, as illustrated in figure 4(e). After separately processing each modality, the fused representations support a more balanced assessment of biases, with both modalities contributing to downstream reasoning (Ghosh et al., 2023; Lei and Huang, 2024, 2025; Baez Santamaria et al., 2024; Preciado Márquez et al., 2025; Panayotov et al., 2022; Manzoor et al., 2025). The fusion between two modalities can occur at an early or late stage.

In early fusion, the graph encoder and text encoder operate as separate components whose outputs are coordinated through a mapping objective or projected into a shared semantic space (Ghosh et al., 2023; Lei and Huang, 2024, 2025; Baez Santamaria et al., 2024; Preciado Márquez et al., 2025). For instance, Lei and Huang (2025) constructs a multi-document event relation graph from ideologically diverse articles. The graph structure is encoded using a GNN, while a textualized graph is simultaneously passed through an LLM. The two representations are then fused within a shared self-attention layer to generate a neutralized summary.

In late fusion, outputs from the graph and text encoders are combined at the final prediction stage without projecting them into a shared semantic space or modifying the internal architecture of LLMs (Manzoor et al., 2025; Panayotov et al., 2022). Manzoor et al. (2025) addresses structural disconnection in graphs by introducing global knowledge into graph representations. The label distributions predicted by the graph (e.g., GNN) are concatenated with those generated by the LLM to obtain the final label distribution.

3.3 Systematic Comparisons

Table 2 shows that a small number of studies are evaluated on overlapping datasets, metrics, and tasks, which limits meaningful comparison across graph-based integration approaches for fairness. For example, graph-retrieval-augmented LLMs are evaluated on HateXplain (Mathew et al., 2021) and IHC (ElSherief et al., 2021), but are compared against only a single work. Similarly, graph-LLM fusion is evaluated on NeUS (Lee et al., 2022), BASIL (Fan et al., 2019), and IHC (ElSherief et al., 2021), yet across different tasks. This highlights the need for standardized evaluation protocols and shared datasets to support generalizable and more reliable assessment graph-based fairness studies.

4 Opportunities and Future Directions

Motivated by gaps in current graph-based fairness studies, we outline potential opportunities and future research directions.

4.1 Framework Specialization for Fairness

Graph-LLM integration has largely focused on improving factual reasoning and verification through structured information. Although factual accuracy and fairness address different challenges, their underlying integration pipelines and architectures are often similar. (Pan et al., 2024). Therefore, we propose two directions to develop more specialized frameworks for fairness applications.

Fairness-sensitive Evaluation Protocols Graph construction introduces multiple sources of uncertainty. Fine-tuning approaches may inherit social biases from relation extraction datasets (Stranisci et al., 2024) and achieve relatively low extraction accuracy (Lei and Huang, 2024, 2025). LLMs can generate hallucinated triplets, while automatic information extraction systems produce noisy data (Deshpande et al., 2022). These challenges call for comprehensive evaluation protocols that combine intrinsic and extrinsic assessments, such as scoring-function analysis and graph-completion benchmarks, alongside human validation.

Moreover, fairness mitigation should preserve factual and biologically grounded biases (e.g., breast cancer is more common in women) rather than enforcing uniform outputs across social groups (Chen et al., 2025). Evaluation frameworks should therefore include metrics that assess whether commonsense and factual knowledge are retained after debiasing, ensuring a balance between fairness and factual integrity.

Explainability of Graphs Graph structures can enhance interpretability by enabling transparent tracing of reasoning paths. Because social bias is often nuanced and context-dependent, future work should prioritize interpretable explanations or counter-speech that clarify how and why bias

Work	Score	Dataset	Metric	Task	Integration Framework
Lei and Huang (2025)	1.26/0.71	Neus	Arousal ↓	Debiased Generation	Fusion, LLM-Driven
Liu et al. (2023a)	6.12/3.60	Neus	Arousal ↓	Debiased Generation	LLM-Driven
Zhao et al. (2025)	72.38	HateXplain	F1-score ↑	Bias Detection	Graph-RAG, LLM-Driven
Wasi (2024)	74.7	HateXplain	F1-score ↑	Reasoning Generation, Bias Detection	Graph-Enhanced
Lei and Huang (2024)	52.00	BASIL	F1-score ↑	Bias Detection	Fusion, LLM-Driven
Liu et al. (2023b)	68.50	BASIL	F1-score ↑	Bias Detection	LLM-Driven, Graph-Enhanced
Ghosh et al. (2023)	64.65	IHC	F1-score ↑	Bias Detection	Fusion
Zhao et al. (2025)	69.95	IHC	F1-score ↑	Bias Detection	Graph-RAG, LLM-Driven
Chen et al. (2025)	49.7/51.3	CrowsPairs	CrowSPairs Score	Preference Comparison	Graph-Enhanced, Graph-Analyzed
Ma et al. (2024)	48.1/49.2	CrowsPairs	CrowSPairs Score	Debiased Generation, Bias Detection	Graph-Enhanced
Chen et al. (2025)	51.2/51.9	StereoSet	StereoSet Score	Preference Comparison	Graph-Enhanced, Graph-Analyzed
Ma et al. (2024)	58.7/55.0	StereoSet	StereoSet Score	Debiased Generation, Bias Detection	Graph-Enhanced

Table 2: Studies using the same datasets and evaluation metrics. Better performance is indicated by lower arousal scores (↓), higher F1 scores (↑), and scores closer to 50 for Crows-Pairs and StereoSet. Arousal is reported separately for Arousal^{+}/Arousal^{-}, CrowS-Pairs Score and StereoSet Scores are reported separately for Gender/Race.

is detected or mitigated. Such transparency can help users understand the underlying reasoning and better trust model outputs.

4.2 Effectiveness of Integration Strategies

Although we review five types of graph-LLM integration, systematic comparisons of which integration stage is most effective for fairness (e.g., fine-tuning, input augmentation, representation learning) remain largely unexplored. Combining multiple integration stages may offer complementary benefits, but their relative and joint contributions are not well understood. Thus, comparative analyses across integration stages are needed to identify where integrations exert the strongest effect and guide future work design.

4.3 Linguistic Diversity and Type-Specific Biases

Most existing studies focus on English datasets. Because the salience and sensitivity of particular biases vary by linguistic and cultural context, bias mitigation strategies effective in English may not generalize to other languages. Expanding fairness studies on multilingual and cross-cultural contexts is thus essential for the robustness of fairness across languages (Ramesh et al., 2023).

In addition, future studies should move beyond broad social bias to examine specific and under-explored bias types such as religion, disability, or age. Besides ideological bias, most existing studies often rely on broad, multi-type social bias datasets rather than specific corpora, creating type imbalance even within a dataset. Targeted studies on individual bias types can enable a deeper understanding of how bias manifests in different contexts and support more precise mitigation strategies.

4.4 LLMs as Agent

With growing reasoning and decision-making capabilities of LLMs, the paradigm of *LLMs as agents* has emerged in recent research, where models autonomously plan, reason, and retrieve information (Ren et al., 2024). Recent frameworks, including RoG (Luo et al., 2024) and ToG (Sun et al., 2024), present this approach by enabling LLM agents to interact with knowledge graphs for question answering. Future research could extend these systems to agentic LLMs that dynamically update and refine the information along with graphs. Such interactive frameworks may improve reasoning transparency, identify sources of biased inference, and correct inaccurate or outdated graph information.

5 Conclusion

This survey reviews research at the intersection of fairness in NLP and graph-based representations. We first introduced the overviews of fairness and graph representations, then presented taxonomies that describe the roles of graphs – structure analysis, external knowledge, and contextual representation– and categorized graph-LLM integration frameworks. Building on identified challenges, we proposed future directions for developing fairness-specialized frameworks, more reliable evaluations, examining effective integration strategies, and agentic LLM systems. These directions highlight the potential of graph-LLM integration to advance fairness in NLP while improving the explainability and trustworthiness of LLMs.

Limitations

We acknowledge several limitations of this survey. First, we do not provide definitions or detailed methods of fairness and graphs. Instead, we offer an overview grounded in prior research and organize the reviewed literature according to defined categories, focusing on graph-LLM integration strategies for fairness rather than coverage of each area. Second, although social bias is often categorized by source, such as representation bias or algorithmic bias, we do not structure our analysis along this dimension. Nevertheless, many reviewed approaches implicitly address different bias sources through mechanisms such as architectural modification or knowledge augmentation. Third, while we outline the roles of graphs in fairness and integration categories, these are not exhaustive and may overlap with general graph-LLM frameworks. Nevertheless, we emphasize fairness-specific perspectives by highlighting how and why graphs contribute to bias detection and mitigation. Finally, the majority of the reviewed literature focuses on English language datasets, which may limit the generalizability, and some recent works may have been omitted. As ethical concerns surrounding LLMs continue to grow, future surveys can build upon and extend this work.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: a nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

Selene Baez Santamaria, Helena Gomez Adorno, and Ilia Markov. 2024. Contextualized graph representations for generating counter-narratives against hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7664–7674, Miami, Florida, USA. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374, Online. Association for Computational Linguistics.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Kurt Bollacker, Robert Cook, and Patrick Tufts. 2007. Freebase: a shared database of structured general human knowledge. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, page 1962–1963. AAAI Press.

Necva Bölücü and Pelin Canbay. 2021. Hate speech and offensive content identification with graph convolutional networks. In *FIRE (Working Notes)*, pages 44–51.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Helena Bonaldi, María Estrella Vallecillo-Rodríguez, Irune Zubiaga, Arturo Montejo-Raez, Aitor Soroa, María-Teresa Martín-Valdivia, Marco Guerini, and Rodrigo Agerri. 2025. The first workshop on multilingual counterspeech generation at COLING 2025: Overview of the shared task. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 92–107, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiahang Cao, Jinyuan Fang, Zaiqiao Meng, and Shangsong Liang. 2024. Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Computing Surveys*, 56(6):1–42.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Ruizhe Chen, Yichen Li, Jianfei Yang, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. 2025. Identifying and mitigating social bias knowledge in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 651–672, Albuquerque, New Mexico. Association for Computational Linguistics.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 67–78, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023. CoSyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6159–6173, Singapore. Association for Computational Linguistics.

Goran Glavaš, Vanja Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NIPS*.

Xuhui Jiang, Chengjin Xu, Yinghan Shen, Xun Sun, Lumingyuan Tang, Saizhuo Wang, Zhongwu Chen, Yuanzhuo Wang, and Jian Guo. 2025. On the evolution of knowledge graphs: A survey and perspective. *Preprint*, arXiv:2310.04835.

Kyohoon Jin, Juhwan Choi, JungMin Yun, Junho Lee, Soojin Jang, and YoungBin Kim. 2025. CoBA: Counterbias text augmentation for mitigating various spurious correlations via semantic triples. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10271–10289, Suzhou, China. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2024. Sentence-level media bias analysis with event relation graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5225–5238, Mexico City, Mexico. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2025. Multi-document summarization through multi-document event relation graph reasoning in LLMs: a case study in framing bias mitigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26603–26619, Vienna, Austria. Association for Computational Linguistics.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024a. A survey on fairness in large language models. *Preprint*, arXiv:2308.10149.

Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2024b. A survey of graph meets large language model: progress and future directions. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI-PRICAI-2020, page 3629–3636. International Joint Conferences on Artificial Intelligence Organization.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Siyi Liu, Hongming Zhang, Hongwei Wang, Kaiqiang Song, Dan Roth, and Dong Yu. 2023a. Open-domain event graph induction for mitigating framing bias. *Preprint*, arXiv:2305.12835.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *Preprint*, arXiv:2308.05374.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Yujian Liu, Xinliang Frederick Zhang, Kaijian Zou, Ruihong Huang, Nick Beauchamp, and Lu Wang. 2023b. All things considered: Detecting partisan events from news media with cross-article comparison. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15472–15488, Singapore. Association for Computational Linguistics.

Chu Fei Luo, Ahmad Ghawanmeh, Kashyap Coimbatore Murali, Bhimshetty Bharat Kumar, Murli Jadhav, Xiaodan Zhu, and Faiza Khan Khattak. 2025. Red-teaming for uncovering societal bias in large language models. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 522–537, Vienna, Austria. Association for Computational Linguistics.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*.

Congda Ma, Tianyu Zhao, and Manabu Okumura. 2024. Debiasing large language models with structured knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10274–10287, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Mandl, Koyel Ghosh, Nishat Raihan, Sandip Modha, Shrey Satapara, Tanishka Gaur, Yaashu Dave, Marcos Zampieri, and Sylvia Jaki. 2025. Overview of the hasoc track 2024: Hate-speech identification in english and bengali. In *Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '24, page 1–2, New York, NY, USA. Association for Computing Machinery.

Muhammad Arslan Manzoor, Ruihong Zeng, Dilshod Azizov, Preslav Nakov, and Shangsong Liang. 2025. MGM: Global understanding of audience overlap graphs for predicting the factuality and the bias of news media. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7279–7295, Albuquerque, New Mexico. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1(1).

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

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

Panayot Panayotov, Utsav Shukla, Husrev Taha Sencar, Mohamed Nabeel, and Preslav Nakov. 2022. GREENER: Graph neural networks for news media profiling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7470–7480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

David Salvador Preciado Márquez, Helena Gómez Adorno, Ilia Markov, and Selene Baez Santamaria. 2025. NLP@IIMAS-CLTL at multilingual counterspeech generation: Combating hate speech using contextualized knowledge graph representations and LLMs. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 29–36, Abu Dhabi, UAE. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh Chawla, and Chao Huang. 2024. A survey of large language models for graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6616–6626. ACM.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, page 148–163, Berlin, Heidelberg. Springer-Verlag.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in

coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

Abel Salinas, Louis Penafiel, Robert McCormack, and Fred Morstatter. 2024. "i'm not racist but...": Discovering bias in the internal knowledge of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marco Stranisci, Pere-Lluís Huguet Cabot, Elisa Bassignana, and Roberto Navigli. 2024. Dissecting biases in relation extraction: A cross-dataset analysis on people's gender and origin. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 190–202, Bangkok, Thailand. Association for Computational Linguistics.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *Preprint*, arXiv:2307.07697.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Denny Vrandečić. 2012. Wikidata: a new platform for collaborative data collection. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 1063–1064, New York, NY, USA. Association for Computing Machinery.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, and 1 others. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Azmine Toushik Wasi. 2024. Explainable identification of hate speech towards islam using graph neural networks. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 250–257, Miami, Florida, USA. Association for Computational Linguistics.

Tian Xie, Tongxin Yin, Vaishakh Keshava, Xueru Zhang, and Siddhartha Reddy Jonnalagadda. 2025. Biascause: Evaluate socially biased causal reasoning of large language models. *Preprint*, arXiv:2504.07997.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing

methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Yibo Zhao, Jiapeng Zhu, Can Xu, Yao Liu, and Xiang Li. 2025. Enhancing LLM-based hatred and toxicity detection with meta-toxic knowledge graph. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24747–24760, Vienna, Austria. Association for Computational Linguistics.

Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. 2023. A comprehensive survey on automatic knowledge graph construction. *ACM Comput. Surv.*, 56(4).

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. 2021. WIKIBIAS: Detecting multi-span subjective biases in language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1799–1814, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

A Appendix

Dataset	Bias Type	Source	Language
WinoBias (Zhao et al., 2018)	Demg.	Synthetic	EN
WinoGender (Rudinger et al., 2018)	Demg.	Synthetic	EN
BASIL (Fan et al., 2019)	Ideog.	News	EN
Reddit (Qian et al., 2019)	Demg., Ident.	Social media	EN
GAB (Qian et al., 2019)	Demg., Ident.	Social media	EN
BiasedSents (Lim et al., 2020)	Ideog.	News	EN
ACL-2020 (Baly et al., 2020)	Ideog.	News	EN
CrowS-Pairs (Nangia et al., 2020)	Demg., Ident.	Crowdsourced	EN
Xhate-999 (Glavaš et al., 2020)	Demg., Ident.	News, Social media, Wikipedia	EN, SQ, HR, DE, RU, TR
SBIC (Sap et al., 2020)	Demg., Ident.	Social media	EN
WikiBias (Zhong et al., 2021)	Demg., Ideog.	Wikipedia	EN
CAD (Vidgen et al., 2021)	Demg., Ident., Ideog.	Social media	EN
ToxicSpans (Pavlopoulos et al., 2021)	Demg., Ident.	Social media	EN
IHC (ElSherief et al., 2021)	Demg., Ident.	Social media	EN
StereoSet (Nadeem et al., 2021)	Demg., Ident.	Crowdsourced	EN
BOLD (Dhamala et al., 2021)	Demg., Ident., Ideog.	Wikipedia	EN
BBQ (Parrish et al., 2022)	Demg., Ident.	Synthetic	EN
HolisticBias (Smith et al., 2022)	Demg., Ident., Ideog.	Synthetic	EN
NeuS (Lee et al., 2022)	Ideog.	News	EN
BIGNEWS (Liu et al., 2022)	Ideog.	News	EN
DIALOCONAN (Bonaldi et al., 2022)	Demg., Ident.	Social media	EN
HateXplain (Mathew et al., 2021)	Demg., Ident.	Social media	EN
HASOC (Mandl et al., 2025)	Demg., Ident., Ideog.	Social media	EN, BN
ML-MTCONAN-KN (Bonaldi et al., 2025)	Demg., Ident.	Crowdsourced, News, Wikipedia	EN, EU, IT, ES
DECODINGTRUST (Wang et al., 2023)	Demg., Ident.	Synthetic	EN
BiasScope (Chen et al., 2025)	Demg., Ident.	Crowdsourced, Synthetic	EN

Table 3: A summary of datasets used in the reviewed literature and the representative datasets in fairness research. Demg.,Ident.,and Ideog. indicate demographic, identity-based, and ideological bias, respectively.

Task	Data Source
Knowledge Base	WordNet (Miller, 1992) ConceptNet (Liu and Singh, 2004) Freebase (Bollacker et al., 2007) DBpedia (Auer et al., 2007) Google Knowledge Graph[1] Wikidata (Vrandečić, 2012)
Event RE	ACE 2005 (Mitchell et al., 2005) NYT (Riedel et al., 2010) RED (O'Gorman et al., 2016) DocRED (Yao et al., 2019) MAVEN (Wang et al., 2020) MAVEN-ERE (Wang et al., 2022)
Coreference	CoNLL-2011 (Pradhan et al., 2011) CoNLL-2012 (Pradhan et al., 2012) Winogrande (Sakaguchi et al., 2019) KnowRef (Emami et al., 2019)
Temporal RE	TimeBank-Dense (Chambers et al., 2014) THYME (Bethard et al., 2015, 2016, 2017) MATRES (Ning et al., 2018)
Causal RE	Causal-TB (Mirza et al., 2014) EventStoryLine (Caselli and Vossen, 2017) CausalBank (Li et al., 2020)

[1] https://developers.google.com/knowledge-graph

Table 4: Data sources commonly used for graph-construction and reasoning tasks. RE refers to relation extraction.

Work	Task	Role	Integration
Bölücü and Canbay (2021)	Bias Detection	Context	Graph-Enhanced
Deshpande et al. (2022)	Knolwedge Detection	External	Graph-Enhanced
Panayotov et al. (2022)	Bias Detection	External	Fusion
Liu et al. (2023a)	Debiased Generation	Structure	LLM-Driven
Liu et al. (2023b)	Bias Detection	Structure	LLM-Driven, Graph-Enhanced
Salinas et al. (2024)	Prompt Generation	Structure	Graph-Analyzed
Ghosh et al. (2023)	Bias Detection	Context	Fusion
Ma et al. (2024)	Debiased Generation, Bias Detection	Structure	Graph-Enhanced
Lei and Huang (2024)	Bias Detection	Structure	Fusion, LLM-Driven
Baez Santamaria et al. (2024)	Debiased Generation	Context	Fusion
Wasi (2024)	Reasoning Generation, Bias Detection	Context	Graph-Enhanced
Lei and Huang (2025)	Debiased Generation	Structure	Fusion, LLM-Driven
Manzoor et al. (2025)	Bias Detection	External	Fusion
Xie et al. (2025)	Reasoning Generation	Structure	Graph-Analyzed
Zhao et al. (2025)	Bias Detection	External	Graph-RAG, LLM-Driven
Luo et al. (2025)	Jailbreaking	External	Graph-RAG, LLM-Driven
Preciado Márquez et al. (2025)	Debiased Generation	Context	Fusion
Chen et al. (2025)	Preference Comparison	Structure	Graph-Enhanced, Graph-Analyzed
Jin et al. (2025)	Bias Detection	Structure	LLM-Driven

Table 5: A summary of fairness research that uses structure representations, categorized by task, role of graph, and integration strategy. An increasing number of such studies have emerged in recent years across diverse tasks.