Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.066351

ARTICLE

Dual-Perspective Evaluation of Knowledge Graphs for Graph-to-Text Generation

Haotong Wang[*,#], **Liyan Wang**[#] **and Yves Lepage**

Graduate School of Information, Production and Systems, Waseda University, Fukuoka, 808-0135, Japan

*Corresponding Author: Haotong Wang. Email: wanghaotong0925@toki.waseda.jp

[#]These authors contributed equally to this work

Received: 28 March 2025; Accepted: 28 April 2025; Published: 09 June 2025

ABSTRACT: Data curation is vital for selecting effective demonstration examples in graph-to-text generation. However, evaluating the quality of Knowledge Graphs (KGs) remains challenging. Prior research exhibits a narrow focus on structural statistics, such as the shortest path length, while the correctness of graphs in representing the associated text is rarely explored. To address this gap, we introduce a dual-perspective evaluation framework for KG-text data, based on the computation of structural adequacy and semantic alignment. From a structural perspective, we propose the Weighted Incremental Edge Method (WIEM) to quantify graph completeness by leveraging agreement between relation models to predict possible edges between entities. WIEM targets to find increments from models on "unseen links", whose presence is inversely proportional to the structural adequacy of the original KG in representing the text. From a semantic perspective, we evaluate how well a KG aligns with the text in capturing the intended meaning. To do so, we instruct a large language model to convert KGs into natural language and measure the similarity between generated and reference texts. Based on these computations, we apply a Top-K union method, integrating the structural and semantic modules, to rank and select high-quality KGs. We evaluate our framework against various approaches for selecting few-shot examples in graph-to-text generation. Experiments on the Association for Computational Linguistics Abstract Graph Dataset (ACL-AGD) and Automatic Content Extraction 05 (ACE05) dataset demonstrate the effectiveness of our approach in distinguishing KG-text data of different qualities, evidenced by the largest performance gap between top- and bottom-ranked examples. We also find that the top examples selected through our dual-perspective framework consistently yield better performance than those selected by traditional measures. These results highlight the importance of data curation in improving graph-to-text generation.

KEYWORDS: Knowledge graph evaluation; graph-to-text generation; scientific abstract; large language model

1 Introduction

Data quality is commonly defined through "fitness for use", emphasizing a practical and user-driven perspective [1]. As a structured form of data, Knowledge Graphs (KGs) embed knowledge into representations of nodes, edges, and graph topology [2], facilitating a wide range of applications such as question answering, knowledge reasoning, and graph-to-text generation [3]. Current research offers various dimensions for evaluating knowledge graph quality, including intrinsic, contextual, representation, and accessibility [4]. However, the problem is how we can select and assess truly appropriate dimensions for a specific downstream task and how such a task-specific evaluation should be conducted. This paper investigates a task-oriented evaluation framework specifically for the graph-to-text generation task.

Graph-to-text generation is a task that transforms structured knowledge graphs into descriptive natural language text [5]. This task imposes strict requirements on the quality of the input knowledge graph. Without complete structure or semantic fit to the source, the generated text risks gaps or inaccuracies. For output quality, the text must deliver coherence and consistency to reflect the source accurately. Coherence ensures a logical flow and semantic connectivity throughout the text [6], while consistency guarantees unambiguous content with sufficient and accurate information [7]. While considerable research has focused on improving generation models, relatively little attention has been paid to the quality of the input knowledge graph. On the input side, coherence and consistency manifest in the graph's ability to exhibit strong connectivity, relational completeness, and precise graphical representation. More specifically, this can be categorized into two aspects: the graph's intrinsic properties [4], including the connectivity and completeness of its edges, which we refer to as structural adequacy, and its extrinsic semantic representation. Because humans cannot directly interpret knowledge graphs, they must be transformed into a more accessible form and then validated; this property we refer to as semantic alignment. Current studies rarely propose unified KG evaluation approaches that address structural adequacy and semantic alignment, especially in generation tasks. The problem lies in how to construct an evaluation framework to quantify structural adequacy and semantic alignment effectively.

Graph connectivity describes the structural cohesion of a graph. In research on Graph Neural Networks (GNNs) [8,9], the adjacency matrix, while entities and relations are represented through knowledge graph embeddings [10]. In graph theory [11], it is quantified through measures such as density and clustering coefficient. These statistical methods assess connectivity by focusing on the existential nature of edges. However, in practice, A) in a knowledge graph, each node is linked to at least one other node to ensure baseline connectivity, and B) given a fixed number of nodes, the number of edges in a knowledge graph is inherently bounded. As a result, such methods are often insensitive to variations in the presence of edges, making it challenging to capture subtle structural differences. This limitation shifts our focus to its counterpart—unseen links, unseen connections that reveal structural gaps. Unseen links assume a graph possesses baseline connectivity but reveal vulnerabilities when previously unrecognized relationships are identified. The presence of multiple unseen links indicates reduced connectivity and insufficient structural completeness, impairing the graph's suitability to support high-quality text generation. Sensitivity to unseen links is closely tied to the task of Knowledge Graph Completeness [12], which evaluates how well entities and relationships represent the target content. However, completeness assessment requires a reference standard, which is often unavailable in graph-text paired datasets due to the absence of a definitive "gold standard". Knowledge graph completion [13] aims to enhance the internal completeness of a knowledge graph by inferring and adding missing entities or relations. To bridge this gap, we propose an Weighted Incremental Edge Method (WIEM) based on unseen links, where a higher number of such links indicates a less cohesive structure and a failure to represent the true nature of the knowledge graph.

To measure semantic alignment, we leverage the advanced comprehension abilities of Large Language Models (LLMs) to transform graphs into text. Previous work [14] has demonstrated that, with well-designed prompts, LLMs can effectively perform zero-shot graph-to-text generation. If the graph accurately conveys its intended semantics, the text generated by the LLM should closely align with the reference text. Similarly, in evaluating structural adequacy, we also utilize LLMs by predicting relationships between nodes based on the given text. However, to more reliably identify unseen links, we adopt the Princeton University Relation Extraction (PURE) [15] system, which is trained for enhanced relation prediction. We combine the outputs of the PURE model and the LLMs to form model consensus for computing the WIEM. To integrate both structural and semantic perspectives, we apply a Top-K union method to select high-quality knowledge graph samples.

This paper presents scientific abstract generation as a concrete example task [9] that requires precise lexical selection and high-level linguistic expression. Generating scientific abstracts imposes strict quality requirements on the input KGs and significantly increases their complexity at the paragraph level. To assess the generalizability of our approach beyond scientific writing, we additionally conduct experiments on the ACE05 dataset, which consists of sentence-level examples from general-purpose text. Therefore, selecting high-quality KGs for model training is complex and demanding. To validate whether the KGs identified by the proposed evaluation framework is "good" or "bad" examples, we mainly employ in-context learning [16] to validate how the selected support examples influence the quality of graph-to-text generation outcomes. In summary, the contributions of this paper are as follows:

- We introduce a novel evaluation framework tailored for graph-to-text generation that assesses structural adequacy and semantic alignment, addressing the gap left by traditional, task-oriented methods.
- We propose Weighted Incremental Edge Method (WIEM), which leverages the consensus between a large language model and a relation extraction system to capture unseen links, providing a more sensitive measure of graph completeness.
- Our experiments on the ACL-AGD and ACE05 dataset reveal the effectiveness of our structural and semantic evaluations, showing that selecting high-quality knowledge graph samples via the proposed dual-perspective approach consistently improves text generation performance in in-context learning settings.

2 Related Work

2.1 Graph-to-Text Generation

Knowledge graphs (KGs) are beneficial for knowledge storage and comprehension, with widespread applications in tasks such as question answering, retrieval-augmented generation, and text generation [3,17]. The acquisition of knowledge graphs has evolved from manual construction to automated methods, with advancements in related technologies greatly reducing retrieval difficulty [2]. This progress has led to an exponential growth of KG data across various domains. Establishing an evaluation system customized to research needs has become essential, particularly for domain-specific KGs that require specialized frameworks [4]. Improving the quality of KG data directly enhances the cognitive capabilities of the models. However, there is currently a lack of research focused on the selection and evaluation of KGs specifically for graph-to-text generation tasks. This study aims to address this gap by examining the output characteristics of such tasks to identify the features of KGs that meet the requirements for high-quality data.

KGs are not inherently interpretable for humans and thus require transformation into a more accessible format, with natural language text being the most fundamental representation. Current research on graph-to-text generation can be broadly categorized into two approaches based on input format: structure-aware [8,9] and serialized input [7,14]. The former typically employs GNNs to model and learn the relationships between nodes and edges, effectively capturing the structural properties of the graph. The latter involves serializing hierarchical data and utilizing the generative capabilities of LLMs. Although serialized input lacks structural awareness, it significantly enhances the quality of the generated text [14]. However, due to the computational complexity of GNNs in handling large and complex KGs, this paper adopts serialized KGs and LLMs for the evaluation framework and experimental validation.

2.2 Knowledge Graph Quality Evaluation

Data quality is typically assessed based on its applicability, emphasizing which data are usable and which are most effective, particularly in the context of specific tasks [18]. Existing research on KG quality

evaluation can be divided into intrinsic and extrinsic [4]: the inherent properties of the knowledge graph and its performance in specific tasks. This study aligns with these two aspects, focusing on structural and semantic perspectives to provide an appropriate assessment.

From a structural perspective, this study focuses on the connectivity and completeness of the graph. Based on the work of [11], we categorized the most frequently used connectivity measures listed and described in Table 1 into two levels: node-level and global-level measures. However, these measures present several problems. First, some are unsuitable for the graph-to-text generation task. For example, centralization [19] is often measured by identifying the most connected node, but in text, frequently used basic terms such as "method" usually dominate, which do not represent the core idea of a paragraph. Second, measures like degree [20] and clustering coefficient [21] measure the number of connections. In KGs, nodes do not exist in isolation, and the number of edges typically starts from a baseline value. Similarly, the Characteristic Path Length [22] is influenced by the minimum average path length determined by node values. Measures with such baseline dependencies are often insensitive to structural variations.

Table 1: Classification of connectivity measures

Measure	Description
	Node-level
Degree [19,20]	Measures the number of connections of a single node.
Betweenness [19]	Measures a node's role as an intermediary in shortest paths.
Closeness [19]	Measures the average shortest path length from a node to all other nodes.
Clustering Coefficient [21]	Measures the proportion of actual connections to possible connections among a node's neighbors.
Disruption Index [21]	Assesses the impact of removing a node on the overall network connectivity.
Centralization (node-level) [19]	Indicates the degree to which a single node is central compared to other nodes in the network.
	Global-level
Characteristic Path Length [22]	Measures the average shortest path length between all node pairs.
Density [20,23]	Measures the ratio of actual edges to the maximum possible edges.
Centralization (network-level) [19]	Measures the prominence of the most central node relative to others.
Weakly Connected Component Ratio [24]	Measures the structural fragmentation of a graph by computing the ratio between the number of weakly connected components and the total number of nodes.

This study proposes an alternative metric to address these limitations by considering the number of unseen edges. This approach differs from conventional statistical methods by incorporating an understanding of the reference text and its graph. A larger number of unseen edges suggests that the original knowledge graph misses more semantic relations, indicating incomplete coverage of the text's meaning. This incremental perspective is more sensitive to variations and provides a nuanced evaluation of the graph's structural adequacy.

The semantic representation of a KG cannot be directly interpreted, making it necessary to convert KGs into textual form, as in graph-to-text generation tasks. We use an LLM to achieve this, leveraging its powerful understanding capabilities through zero-shot learning. The evaluation focuses on the similarity between the generated text and the reference text. In graph-to-text generation tasks, semantic similarity metrics have gained increasing attention due to their ability to capture meaning beyond surface-level overlap. These metrics, such as BERTScore [25] and BLEURT [26], leverage contextualized embeddings from pretrained language models to assess semantic alignment between the generated and reference texts, offering a more reliable evaluation of fluency, relevance, and informativeness.

3 Methodology

3.1 Overview of the Framework

We propose an evaluation framework tailored to assess the quality of KGs in the graph-to-text generation task. The framework supports data curation through two core dimensions concerning graph structure and its alignment to the associated text. In particular, **structural adequacy** measures whether a graph contains sufficient and coherent relational structure, and while **semantic alignment** evaluates the extent to which the graph aligns with the intended meaning of the target text.

As illustrated in Fig. 1, the evaluation begins with an input knowledge graph G and proceeds through two parallel but complementary modules. The structural evaluation module identifies unseen edges that are semantically justified. To do so, we intersect relational predictions from an LLM and a relation extraction model (PURE). These differences are quantified using Weighted Incremental Edge Method (WIEM), a metric that uses consensus-based weighting driven by model agreement on the likelihood of relational links. In parallel, the semantic evaluation module employs an LLM to generate a textual description of the graph, which is then compared with the reference sentence using BLEURT. BLEURT is a reference-based metric that uses a fine-tuned language model to assess the semantic similarity between texts, achieving high correlation with human evaluation. This dual-perspective approach enables a fine-grained and interpretable evaluation of graph quality, offering actionable guidance for KG selection and sampling in the downstream generation task.

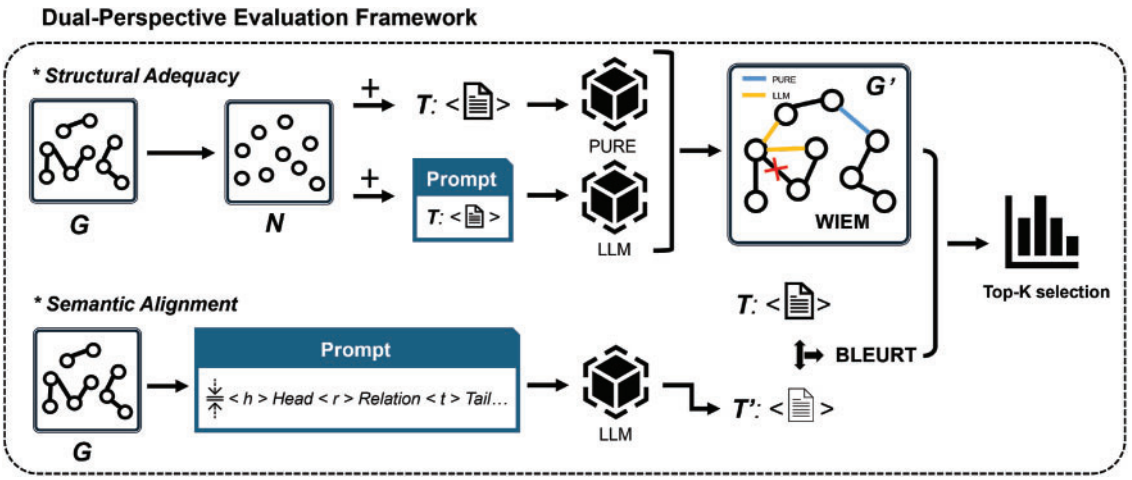

Figure 1: Overview of the dual-perspective evaluation framework

3.2 Structural Adequacy

From a cognitive perspective, a well-structured knowledge graph should exhibit explicit connectivity and align with human expectations regarding implicit semantic relationships. Humans often interpret meaning not solely through direct links but also via latent associations such as co-occurrence, causality, and abstraction. However, due to the limitations of current KG construction models, many of these implicit relations are not encoded, resulting in sparse graphs. The absence of such latent edges undermines the structural completeness of the graph concerning its intended semantics. In practical KG-to-text settings, graphs rarely come with complete annotation. Without a gold standard, we propose using model-generated consensus relations, e.g., from LLMs and relation extractors, as a form of soft ground truth. This approximation enables the estimation of latent structure gaps without human annotation. By synthesizing predictions from high-capacity models, we simulate a form of weak supervision to uncover potential unseen edges.

Given a set of identified entities N and the corresponding reference text T, we employ two different models (PURE and an LLM) to infer potential relational edges (as shown in Fig. 1). PURE is a relation extraction system composed of independent entities and relation encoders. that identifies entities in text and predicts relations between entity nodes. Since the entity set N is already provided in our setting, we fine-tune only the relation extraction module of PURE on the SciERC dataset [27], which shares domain similarity with our target scientific texts. The LLM model is Llama3, which is prompted using both N and T to perform open-ended relational inference. By comparing the predicted edges from these models with those present in the original graph G, we compute edge increments. These inferred edges are then aggregated and weighted using our proposed WIEM to quantify the structural adequacy of the augmented graph G'. The formal definition is provided in the following sections.

3.2.1 Incremental Edge Method

Assuming e represents the number of edges and n is the number of nodes in a graph G, the minimum number of edges required to ensure each node is connected to at least one other node is given by: $e_{\min} = \left\lceil \frac{n}{2} \right\rceil$ the maximum number of edges corresponds to a fully connected graph: $e_{\max} = \frac{n(n-1)}{2}$.

Let i denote the increment value of unseen edges, referring to the ones present in the new graph G' that were not present in the original graph G, as illustrated by the yellow and blue links in graph G' of Fig. 1. Our analysis focuses on newly inferred edges; we do not consider the reverse scenario where existing edges in G are absent from G', such as the red mark shown in Fig. 1. We further define the edge increment ratio as:

$$0 \leq \frac{i}{e} \leq \frac{e_{\max} - e}{e}, \forall e \in \left[\frac{n}{2}, \frac{n(n-1)}{2} \right] \tag{1}$$

Compared to directly using the absolute increment i, the ratio $\frac{i}{e}$ is less affected by the overall graph size. If $e_{\min}$ represents a sparse graph where e is small, an increment i can have a significant impact on the graph structure, and $\frac{i}{e}$ appropriately increases the weight of the increment. Conversely, if $e_{\max}$ represents a dense graph where e is large, the impact of a single edge increment becomes relatively minor, and $\frac{i}{e}$ balances this discrepancy. Thus, $\frac{i}{e}$ adapts to the characteristics of both sparse and dense graphs, avoiding the imbalance that arises from using either the absolute increment i or the absolute number of edges e alone.

To enhance interpretability, we adjusted the metric so that higher values indicate better structural adequacy. By applying an inverse proportional transformation to $\frac{i}{e}$, the incremental edge method is defined

as:

$$\text{IEM} = \frac{1}{1 + \frac{i}{e}} = \frac{e}{e + i} \tag{2}$$

The range of the IEM is $(0, 1]$. A higher IEM value indicates a graph with better structural adequacy, while a lower value suggests poorer structural quality.

3.2.2 Weighted IEM

While the original IEM assumes that all inferred edges contribute equally to structural incompleteness, this may not reflect the varying reliability of different predictions. In our framework, each inferred edges originates from either the LLM, PURE, or both. Because neither model provides probabilistic confidence scores, we adopt a consensus-based weighting method. Specifically, we define the weighted increment as:

$$i_w = |E_{\text{both}}| + \alpha \left|E_{\text{only}}\right| \tag{3}$$

where E_{both} denotes the set of inferred edges predicted by both models and E_{only} denotes edges predicted by only one of the two models. The parameter $\alpha \in [0, 1]$ controls the relative trust placed in partially agreed edges. We empirically determine $\alpha = 0.75$ through grid search on a test set. A value that is too high may overtrust speculative model predictions and overlook the variance between models, introducing noisy edges. In contrast, a low value may exclude valid relations that only one model predicts. We update the formulation as shown in Eq. (4).

$$\text{WIEM} = \frac{e}{e + i_w} = \frac{e}{e + |E_{\text{both}}| + \alpha \left|E_{\text{only}}\right|} \tag{4}$$

The design of WIEM enables a flexible trade-off between precision and recall in structural completeness estimation while preserving the interpretability and scale invariance of the original IEM.

Toy Example: To illustrate how WIEM handles newly inferred edges, consider a toy graph G with four nodes $\{A, B, C, D\}$ and edges $\{(A, B), (B, C)\}$. Suppose PURE predicts two additional edges $\{(A, C), (C, D)\}$, while the LLM predicts three edges $\{(A, C), (B, D), (C, D)\}$. We then obtain $E_{\text{both}} = \{(A, C), (C, D)\}$ and $E_{\text{only}} = \{(B, D)\}$. If $\alpha = 0.75$, the weighted increment becomes $i_w = |E_{\text{both}}| + \alpha \cdot |E_{\text{only}}| = 2 + 0.75 \times 1 = 2.75$. Substituting $e = 2$ into the WIEM formulation $\text{WIEM} = \frac{e}{e+i_w} = \frac{2}{2+2.75} \approx 0.42$. This relatively low value indicates that the original graph$G$ lacks significant structural completeness compared to the newly inferred edges.

3.3 Semantic Alignment

To evaluate the semantic alignment of a KG, we measure how well it conveys the intended meaning of its corresponding reference text. Rather than evaluating the structure of KGs alone, we leverage a graph-to-text generation setup, using an LLM to generate text from KGs. The core assumption is that if a graph contains sufficient and accurate information, the LLM should be able to produce a text that is semantically aligned with the human-authored reference. It is important to note that the LLM here is not the evaluation target but a proxy inference tool. This constitutes a reverse inference process.

For semantic comparison, we employ BLEURT, a reference-based evaluation metric that leverages a pretrained language model fine-tuned on human-annotated ratings. BLEURT is designed to capture subtle semantic differences between the generated and reference texts, and has been shown to correlate strongly with human judgment on fluency, adequacy, and meaning preservation. BLEURT captures both surface

and deep semantic differences, including subtle omissions, hallucinations, or distortions. Compared with surface-level metrics such as BLEU or ROUGE, BLEURT offers a better reflection of semantic consistency in open-generation tasks.

While using LLMs to generate textual descriptions from KGs offers an effective proxy for semantic evaluation, a known risk is their tendency to hallucinate by introducing facts that are not grounded in the input KG. Such hallucinations can artificially inflate BLEURT scores if the generated text aligns with the reference through external knowledge rather than faithful KG representation. To mitigate this risk, we adopt several strategies. First, we fix the generation temperature to a low value (0.2) and apply nucleus sampling with a conservative top-p of 0.8, reducing randomness and encouraging the model to focus on the given KG triples. Second, we employ a strict and explicit prompting format (Fig. 2, right) that clearly instructs the model to base the output solely on the provided knowledge graph, without introducing any external assumptions. Finally, by standardizing the prompt template and decoding parameters across all evaluations, we ensure that any hallucinated content, if present, remains minimal and uniformly distributed, preserving fair comparisons between different KG samples.

Relation Inference Prompt

You are an expert in relation extraction for scientific texts. Your goal is to identify all plausible relationships among a given set of entities, based on the provided reference text. However, the only allowable relation types are: <Compare> <Used-for> <Part-of> <Feature-of> <Hyponym-of> <Evaluate-for> <Conjunction>
Context: I have a set of entities and a corresponding reference text. Please identify all meaningful relationships among these entities, treating the reference text as the primary evidence.
Entities (N): [Entity1, Entity2, Entity3, …]
Reference Text (T): "…[the scientific abstract]…"
Task:
(1). List each relationship you can infer between the entities in (N).
(2). Present each relationship in the format (EntityA, RelationType, EntityB).
(3). If no valid relation type applies, do not invent your own.
(4). If multiple relations exist, list them separately.
(5). Exclude any new entities not found in (N).
Example Output:
(EntityA, Used-for, EntityB)
(EntityC, Compare, EntityD)
…

KG-to-Text Generation Prompt

You are an expert science writer. Your goal is to transform structured knowledge graph data into coherent, factual scientific text.
Context:
I have a knowledge graph (KG) represented by a list of edges in the following format:
< h > Head < r > Relation < t > Tail .
Please generate a concise paragraph that accurately reflects these edges. Use formal academic language, and do not introduce any information beyond what is explicitly stated in the KG.
Knowledge Graphs:
< h > EntityA < r > Part-of < t > EntityB
< h > EntityB < r > Evaluate-for < t > EntityC
...
Task:
(1). Write a coherent paragraph of no more than 250 words, detailing the relationships indicated by the KG.
(2). Keep the text factual and aligned strictly with the given edges.
(3). Maintain a logical flow and clarity.
(4). Avoid adding new entities or relations not in the KG.
Example Output:
"EntityA is part of EntityB, which is evaluated for EntityC. This highlights ..."

Figure 2: Prompts used in the structural and semantic evaluation modules. The left prompt guides relation inference from reference text and entities. The right prompt generates text from KG triples for semantic comparison

3.4 Top-K Union Selection Method

To determine the overall quality of a knowledge graph, we adopt a Top-K union method that combines both structural and semantic evaluations. This method avoids relying on hard thresholding by selecting the Top-K of graphs from each individual ranking, one based on WIEM and the other on BLEURT, and considering their union as "Good" samples.

4 Experiments

4.1 Experimental Settings

To evaluate our framework, we compare it to various data curation approaches in the knowledge graph-to-text generation task. Given a set of KG-text pairs, each method is applied to select a subset of examples, which are then used as in-context demonstrations to prompt LLMs to generate text from a test graph.

Dataset

Experiments are conducted on the ACL Abstract Graph Dataset (ACL-AGD)[1] [28], consisting of 46,282 KG-text pairs. Each pair includes a scientific abstract in the Natural Language Processing (NLP) domain and its corresponding graph, with data collected from the ACL Anthology across publications from 1965 to May 2024. We randomly sample 1000 KG–text pairs from the ACL-AGD corpus as a test set for graph-to-text generation. We sample 10,000 instances from the remaining pairs as a candidate pool, to represent the overall distribution while ensuring efficient evaluation across all selection approaches. Each approach then selects the top and bottom 300 pairs from 10,000 candidates based on their ranking measures.

To further validate the generalizability of our proposed method, we additionally employ the ACE05 dataset[2] , which covers a wide range of sources, including broadcast conversations, broadcast news, and newsgroups. In contrast to ACL-AGD, which operates at the paragraph level, ACE05 focuses on sentence-level KG. We follow the preprocessing style of DyGIE[3] to construct the input graphs. PURE [15] provides pretrained entity and relation models on ACE05, we directly utilize the released model parameters. As these models have already been trained on the ACE05 training set, we are unable to perform KG selection on the training portion. Therefore, we restrict our selection and evaluation to the test set, which contains approximately 2050 instances.

Implementation details

For KG-to-text generation, we use the LLaMA3-8B[4] [29] model under the Ollama framework on two Nvidia RTX A6000 GPUs. We set the temperature to 0.2, apply nucleus sampling with top_p of 0.9, and limit the output to 150 tokens. The model is instructed using a prompt template as shown in Fig. 2. We expect the model to generate an abstract-style paragraph for a given KG input, conditional on verbalized instructions on the task and the triplet structure of graphs. For each test instance, a few (one or two) in-context examples are randomly drawn from either the top-300 or bottom-300 set as in-context examples incorporated into the input prompt.

Metrics

We measure the quality of the generated text by comparing it against reference texts using both formal and semantic similarity metrics. For formal similarity, we use ROUGE-L [30], as implemented in HuggingFace's *evaluate* library, to measure the longest common subsequence between texts. For semantic similarity, we adopt BLEURT (BLEURT-20-D6) [26], a pretrained metric fine-tuned on human judgments to capture subtle semantic differences. All experiments are conducted under a fixed random seed to ensure consistency.

[1] The ACL-AGD corpus is publicly available at http://lepage-lab.ips.waseda.ac.jp/projects/scientific-writing-aid (accessed on 28 April 2025).
[2] https://catalog.ldc.upenn.edu/LDC2006T06 (accessed on 28 April 2025).
[3] https://github.com/luanyi/DyGIE/tree/master/preprocessing (accessed on 28 April 2025).
[4] https://ollama.com/library/llama3:8b (accessed on 28 April 2025).

Baselines

Our baselines include the node-level measures Degree, Closeness, the global-level measures Density, Clustering Coefficient (ClustC) and Weakly Connected Component Ratio (wccR). In addition, we evaluate WIEM, Random, Direct LLM, and the proposed method, providing a comprehensive set of structural and semantic measures for comparing KG quality in graph-to-text scenarios:

- **Degree [19]:** This counts the edges tied to one node in a knowledge graph. A higher degree indicates denser information distribution and a lower degree reflects reduced semantic connectivity among entities.
- **Closeness [19]:** It measures a node's average shortest path distance to all other nodes. Shorter distances indicate higher centrality and imply stronger interconnections, reducing scattered information.
- **Density [23]:** The ratio of a graph's actual edges to the maximum possible edges. A higher density indicates a structure closer to a fully connected graph, while a lower value means a sparsity graph.
- **Clustering Coefficient (ClustC) [21]:** It measures how many connections exist among a node's neighbors compared to all possible connections. A higher value shows a tightly knit, cohesive knowledge graph, while a lower value indicates a sparser, less connected structure.
- **Weakly Connected Component Ratio (wccR) [24]:** It quantifies graph fragmentation as the ratio of weakly connected components to total nodes, where higher values indicate many small disconnected subgraphs, and lower values reflect stronger overall connectivity.
- **WIEM:** It focuses on potential unseen edges relative to a reference text. Unlike the statistics methods, which rely on an existing graph, WIEM is more sensitive to whether the KG adequately covers its source semantics.
- **Random:** Randomly selects samples without structural or semantic filtering as a baseline to compare against "unsorted" selection.
- **Direct LLM:** This ranks KGs based on texts generated by an LLM using BLEURT, evaluating how selection driven by text similarity affects generation quality. The proposed method combines WIEM and Direct LLM, a joint approach that considers both structure and semantics.

4.2 Main Results

Table 2 presents the performance comparison under a 1-shot in-context generation setting on the ACL-AGD, where different KG selection methods are evaluated based on their impact on downstream text generation quality. Traditional structure-based methods exhibit varied effectiveness. Closeness shows almost no improvement, with negligible differences between top and bottom selected samples. In contrast, Degree and Density show moderate positive gains. The top-selected samples by Degree outperform the bottom-selected group by 1.58 in ROUGE-L and 2.54 in BLEURT, while Density shows gains of 0.96 and 2.06, respectively. ClustC, which measures the tendency of nodes to form local clusters, yields only marginal improvements of 0.27 in ROUGE-L and 0.31 in BLEURT, indicating limited impact on generation quality. wccR, which reflects the degree of graph fragmentation, achieves slightly higher gains of 0.25 in ROUGE-L and 1.57 in BLEURT, suggesting that graphs with fewer disconnected components tend to support more coherent and semantically aligned text. Our proposed WIEM, which identifies unseen relations through model consensus, consistently delivers stronger improvements, with gains of 1.55 in ROUGE-L and 2.81 in BLEURT. Compared to Closeness, ClustC, and wccR, WIEM is more sensitive to structural incompleteness and proves more effective in identifying higher-quality graphs for generation tasks.

Compared to structural methods, methods incorporating semantic alignment exhibit stronger performance in distinguishing KG-text data of varying quality. The Direct LLM method achieves the second-best results, with a Difference of +3.85 in ROUGE-L and +2.81 in BLEURT. Our proposed method outperforms

all baselines, achieving +5.33 in ROUGE-L and +3.21 in BLEURT. It turns out that blending connectivity and semantic coverage gives us the best shot at selecting high-quality KGs.

Table 2: Effect of knowledge graph selection on 1-shot prompted text generation over the ACL-AGD. Bold numbers indicate the best result in each column

Selection method	ROUGE-L			BLEURT		
	Top ↑	Bottom ↓	Difference	Top ↑	Bottom ↓	Difference
Degree [19]	26.38	24.80	+1.58	43.06	**40.52**	+2.54
Closeness [19]	23.90	**24.03**	−0.13	41.46	41.56	−0.10
Density [23]	25.64	24.68	+0.96	42.94	40.88	+2.06
ClustC [21]	24.92	24.65	+0.27	41.87	41.56	+0.31
wccR [24]	25.14	24.89	+0.25	42.77	41.20	+1.57
Random	23.68	24.19	−0.51	41.65	41.37	+0.28
WIEM	26.16	24.61	+1.55	43.59	40.78	+2.81
Direct LLM	29.45	25.60	+3.85	43.92	41.11	+2.81
Ours	**29.49**	24.16	**+5.33**	**44.12**	40.91	**+3.21**

On a curious note, we observe that bottom-ranked examples do not reduce generation quality. Their performance is often comparable to randomly selected graphs, suggesting that large language models exhibit robustness and can compensate for noisy graph inputs.

Fig. 3 compares generation quality under 0-shot, 1-shot, and 2-shot settings, using support examples selected from the top-selected KGs of each method. The 0-shot setting refers to direct generation on the test set without any support examples (BLEURT = 41.45, ROUGE-L = 23.96). Closeness exhibits minimal and unstable improvements across shot settings, highlighting its ineffectiveness in identifying generation-relevant graph features. Similarly, ClustC and wccR show marginal performance gains, suggesting that clustering tendency and weak connectivity alone are insufficient for selecting high-impact support graphs. Degree and Density demonstrate more consistent improvements, though their relatively shallow gains reflect limited depth in semantic or structural guidance. WIEM and Direct LLM both yield strong 1-shot performance and show only minor variation in the 2-shot setting, indicating that the added value of additional support examples plateaus when selection is already high-quality. Our proposed method consistently achieves the best results across all settings, with BLEURT reaching 44.73 and ROUGE-L reaching 29.51 in the 2-shot case. These results confirm the effectiveness of our dual-perspective evaluation framework in amplifying the utility of selected graphs for in-context generation.

Table 3 presents the experimental results of our 1-shot prompted generation setting conducted on the ACE05 dataset. Compared with Table 2, which is based on paragraph-level knowledge graphs from ACL-AGD, the sentence-level nature of ACE05 leads to generally higher ROUGE-L and BLEURT scores across all methods, owing to the reduced complexity and smaller size of the input graphs. Consequently, the performance gaps between top- and bottom-ranked KGs become narrower. Despite this, our proposed method still achieves the best overall performance, with the largest gains in both ROUGE-L (+2.20) and BLEURT (+2.01). Notably, WIEM also shows strong effectiveness under this setting, ranking second with a ROUGE-L gain of +2.01 and a BLEURT gain of +1.65, significantly outperforming random selection and traditional graph connectivity baselines such as Degree and Closeness. These results demonstrate that our method remains effective even when applied to a completely different dataset with distinct domain characteristics and graph granularity.

Figure 3: Performance comparison of 1-shot and 2-shot prompting on BLEURT and ROUGE-L, with supporting examples sourced from top-selected graphs

Table 3: Effect of knowledge graph selection on 1-shot prompted text generation over the ACE05. Bold numbers indicate the best result in each column

Selection method	ROUGE-L			BLEURT		
	Top ↑	Bottom ↓	Difference	Top ↑	Bottom ↓	Difference
Degree [19]	64.12	63.47	+0.65	56.11	55.50	+0.61
Closeness [19]	63.16	63.79	−0.63	55.61	54.94	+0.67
Density [23]	64.40	63.58	+0.82	56.01	55.46	+0.55
ClustC [21]	63.66	63.80	−0.14	55.49	55.78	−0.29
wccR [24]	65.10	64.03	+1.07	56.03	55.30	+0.73
Random	63.90	63.48	+0.42	55.23	55.77	−0.54
WIEM	65.18	**63.17**	+2.01	56.43	**54.78**	+1.65
Direct LLM	64.89	63.72	+1.17	56.12	55.13	+0.99
Ours	**65.43**	63.23	**+2.20**	**56.80**	54.79	**+2.01**

4.3 Validating Module Effectiveness via Random Edge Perturbation

We assess how structural adequacy (measured by WIEM) and semantic alignment (measured by BLEURT) respond to varying levels of random perturbation applied to knowledge graphs from the ACL-AGD dataset. Specifically, we conduct random edge removal and edge addition at five levels: 10%, 20%, 30%, 40%, and 50%. The original setting refers to the unaltered test set as the baseline for comparison. The motivation is to test the sensitivity of our evaluation framework. By simulating edge-level corruption or redundancy, we can observe how structural disruptions in the knowledge graph affect its ability to support high-quality text generation.

As shown in Fig. 4, random edge removal causes a substantial decline in both WIEM and BLEURT. As the perturbation level increases, WIEM decreases evidently, indicating that more edges are being inferred to compensate for unseen edges. BLEURT also drops, reflecting the loss of important content connections that weaken the alignment between the generated and reference texts. Random edge addition results in much more stable trends: WIEM gradually increases as redundant edges slightly enhance perceived graph connectivity, and BLEURT remains consistent, pointing to these extra edges introducing only minor semantic noise. These results validate that our evaluation framework is effective in distinguishing data quality in KG-to-text generation. Retaining essential relational content ensures logical organization and meaning preservation in graph-to-text generation.

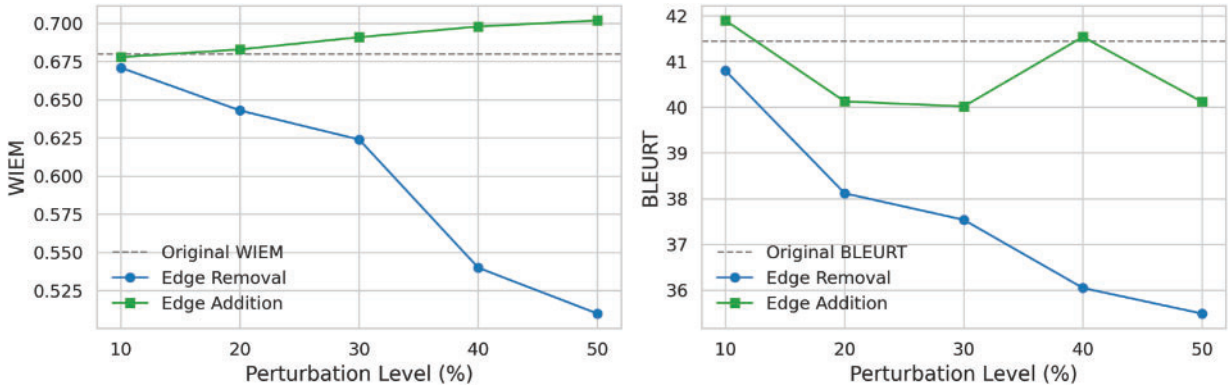

Figure 4: Impact of random edge perturbation on WIEM and BLEURT

Fig. 5 illustrates the positive correlation between WIEM and BLEURT. As BLEURT increases, the WIEM also rises, suggesting that complete graphs align better with meaning. This correlation confirms the value of our evaluation framework, showing how structure ties closely to a graph's meaning.

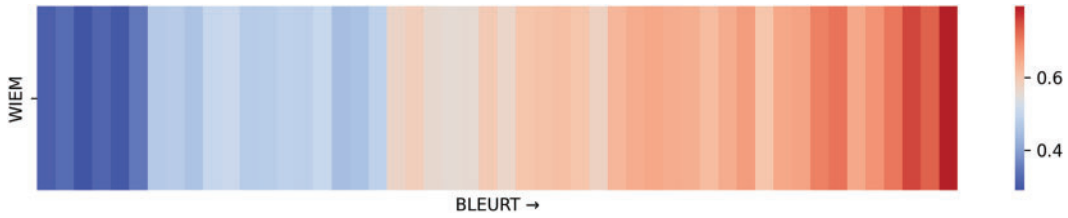

Figure 5: Correlation between WIEM and BLEURT

4.4 Computational Requirements

All experiments were conducted on a workstation configured with an Intel i9-10980XE CPU, 128 GB of RAM, and two NVIDIA RTX A6000 GPUs (each with 48 GB of memory). As shown in Table 4, we report the per-graph inference time, average memory usage, and average GPU memory usage for key baseline methods and major components of our framework. The evaluation is divided into two categories: traditional statistical computations and LLM-based components. The statistical baselines (ClustC and wccR) have minimal inference time (under 0.24 s per graph) and negligible memory usage without relying on GPU resources. In contrast, our framework includes components such as LLM Prompting, which performs both graph reasoning and text generation using the LLaMA3-8B model, with reported values reflecting their

shared usage. PURE and WIEM involve lightweight model inference and structural scoring, respectively, contributing moderate overhead. The Total (Combined) row summarizes the end-to-end computation time and memory usage per graph across the full pipeline. Traditional statistical baselines exhibit minimal computational cost, whereas LLM-based components, especially the prompting stage, introduce substantial GPU and memory demands. Despite this increase, the overall framework maintains reasonable inference time and resource usage, indicating its practicality for large-scale deployment.

Table 4: Per-graph inference time and memory usage for baselines and core components of our framework

Method	Times (s)	Avg memory (GB)	Avg GPU memory (GB)
ClustC [21]	0.146	0.019	–
wccR [24]	0.238	0.021	–
PURE	0.323	0.474	3.432
LLM Prompting	1.241	1.768	25.285
WIEM	0.185	0.038	–
Total (Combined)	3.144	3.741	42.797

Our framework incurs higher computational overhead due to the integration of model-based reasoning and text generation. To improve efficiency, we parallelize WIEM scoring and BLEURT evaluation by processing multiple graphs concurrently where system resources allow. However, LLM Prompting is performed individually for each graph, as the generation process depends on graph-specific prompts and cannot be batched without altering output quality. This partial parallelization helps reduce latency for structural and semantic scoring, although generation remains sequential. Future work may explore more efficient decoding techniques or lightweight model distillation to further accelerate the prompting stage.

4.5 Exploration of Alternative Ranking Methods

To further assess the robustness of our Top-K union selection method, we compare it with two alternative fusion methods: (i) a weighted sum of normalized WIEM and BLEURT ranks, and (ii) the intersection of top-K graphs from each ranking. For Weighted Sum Rank, each graph receives a combined score: $\text{Score}_i = \alpha \cdot \text{rank}_{\text{WIEM}}(i) + (1-\alpha) \cdot \text{rank}_{\text{BLEURT}}$, where $\alpha = 0.5$, reflecting equal contribution from structural and semantic perspectives. Graphs are then selected according to their combined scores. In the intersection approach, only graphs that simultaneously appear in both WIEM and BLEURT top-K rankings are selected.

Using each selected graph set, we conduct one-shot graph-to-text generation and evaluate the outputs using BLEURT and ROUGE-L. As shown in Table 5, the Top-K union method achieves the highest BLEURT (44.12) and ROUGE-L (29.49) scores, confirming its robustness. Compared to the intersection method, which slightly improves BLEURT (43.95) at the cost of reduced sample size and diversity, Top-K union maintains broader coverage and avoids over-pruning. Although Weighted Sum Rank also retains 300 graphs, its performance is less stable (BLEURT = 43.46 ± 0.28), likely due to its sensitivity to the α parameter. These findings demonstrate that Top-K union effectively balances structural and semantic perspectives without sacrificing diversity or stability, making it a reliable fusion strategy for high-quality generation.

Table 5: Comparison of different fusion methods based on sample size and generation quality

Fusion method	Sample size	BLEURT	ROUGE-L
Weighted Sum Rank	=300	43.46 ± 0.28	29.10 ± 0.20
Top-K Intersection	≤300	43.95 ± 0.16	29.22 ± 0.13
Top-K Union ✓	≈300	44.12 ± 0.21	29.49 ± 0.22

4.6 Ablation Studies

4.6.1 Effects of Model Agreement on Edge Weighting

Fig. 6 (left part) illustrates the distribution of WIEM scores under three settings: PURE, which uses only relation extraction results from the PURE model; LLM, which uses only inferred edges from an LLM; and PURE + LLM, which combines predictions from both models to compute WIEM. For the single model, we just calculate IEM.

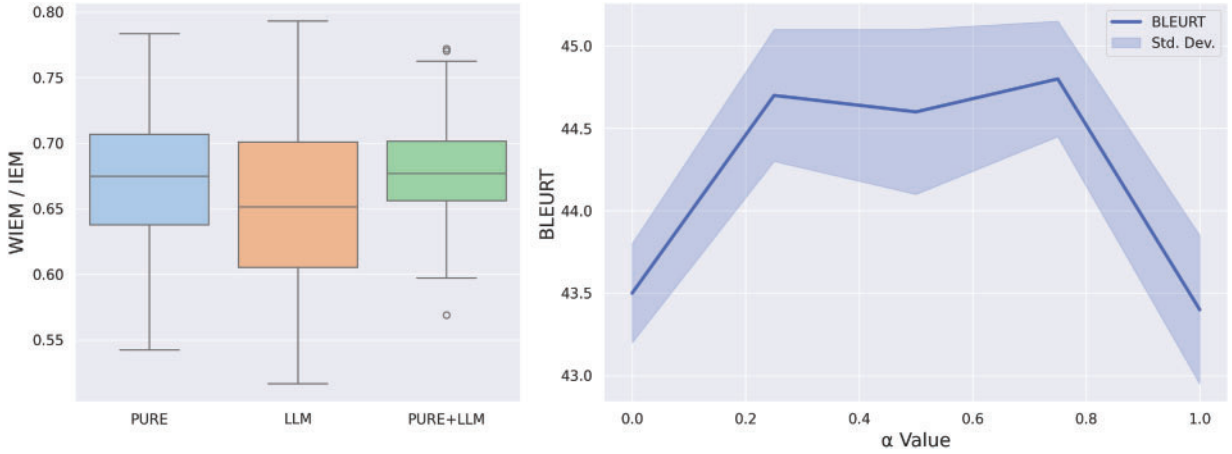

Figure 6: Comparisons of WIEM outcomes under single-model (PURE or LLM) vs. consensus approaches (left) and model performance in graph-to-text generation using one-shot prompts with examples selected under various α settings (right)

WIEM scores tend to be higher and more tightly clustered in the PURE setting, pointing to restrained but solid predictions. The LLM setup has a lower median WIEM but swings wider in variance, showing that LLMs might dig up more potential edges, though their predictions can waver. The combined PURE + LLM setting strikes a balance between the two, capturing a broader range of edges while avoiding noise, achieving a favorable trade-off in both mean score and stability.

4.6.2 Ablation on α Settings

The right part of Fig. 6 shows how BLEURT scores and their standard deviations change as the parameter α increases from 0.0 to 1.0. We perform the graph-to-text generation task using the top-300 knowledge graphs selected under each α value. For each setting, we report the average BLEURT score over 5 independent runs, where the model generates text using one-shot prompting with a different support example sampled from the top-300 set in each run. Here, α controls the relative weight assigned to edges

predicted by only one model vs. those confirmed by both. An α value of 0 discounts single-model predictions, prioritizing consensus, while $\alpha = 1$ treats all predicted edges equally, potentially increasing edge richness at the cost of noise.

The results reveal that performance peaks when α lies in the mid-range, particularly at 0.75, while boundary values (i.e., $\alpha = 0.0$ or 1.0) lead to lower BLEURT scores or increased variability. This observation highlights the need to balance unique predictions and consensus edges: a low α may miss informative edges flagged by a single model, harming completeness, whereas a high α may admit noisy or irrelevant edges, compromising graph quality.

To further support the selection of $\alpha = 0.75$, we compute 95% confidence intervals based on the five BLEURT scores per α setting. The results show that $\alpha = 0.75$ yields the highest mean BLEURT (44.8), with a confidence interval of [44.49, 45.11], clearly outperforming the extremes and remaining competitive with neighboring values like 0.25 and 0.5. This suggests that our framework is robust to moderate shifts in α, and that $\alpha = 0.75$ achieves an effective trade-off between informativeness and noise control. Therefore, the choice of α is empirically grounded in both performance trends and statistical confidence.

4.6.3 Effect of Selected K on Generation Performance

Fig. 7 illustrates the one-shot generation performance across different values of K, representing the number of selected knowledge graphs. A smaller K yields a candidate pool with a higher proportion of top-ranked, high-quality graphs, resulting in purer and more reliable support examples. This setup often leads to higher generation precision. However, a smaller pool also limits the diversity of support examples, which may reduce the generalizability of the model to unseen structures. As K increases, the candidate pool covers a wider range of graph patterns, but it also introduces greater uncertainty in quality. Low-quality graphs may be included, leading to noisier support examples and a degradation in generation performance. This trend is evident in both BLEURT and ROUGE-L scores, which decline when K exceeds 300. To strike a balance between generalization capability and generation quality, we select $K = 300$ as an optimal trade-off, achieving consistently strong performance without compromising stability.

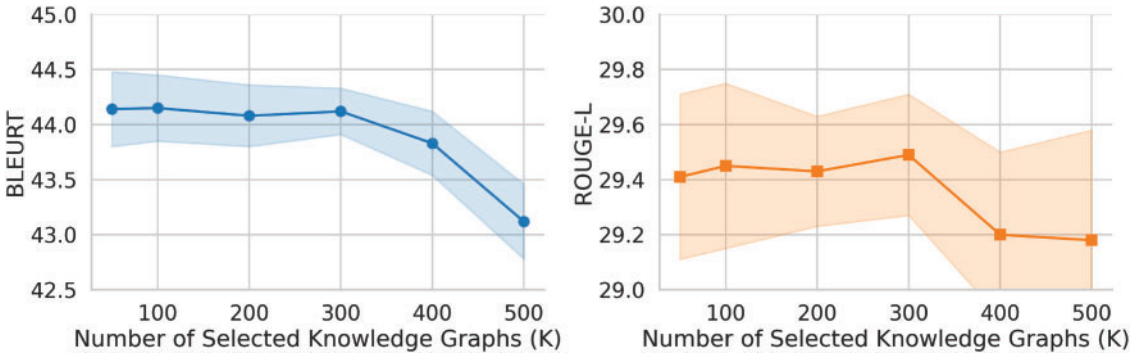

Figure 7: Generation performance under varying selected pool sizes K

4.7 Case Study

We present four knowledge graphs from the top 300 ranked candidates. To avoid cherry-picking, we apply a fixed random seed to uniformly sample four graphs from this subset. This approach ensures fair and unbiased selection, while also highlighting the robustness of our evaluation framework in consistently surfacing high-quality graphs without manual intervention. As shown in Fig. 8, these knowledge graphs demonstrate strong structural connectivity. Even with a few stray links, the structure stays packed with

connections. These chains typically encode specialized or context-specific information rather than indicating fragmented or incomplete knowledge. We skipped the bottom examples since most have barely any links or just standalone ones.

We also observed a close alignment between graph connectivity and semantic richness. Upon qualitative inspection of the sampled graphs, we found that densely connected nodes tend to correspond to key domain-specific concepts (e.g., "dataset", "Twitter") that carry the core semantic load of the reference text. These central entities are typically embedded in well-formed relational chains, enabling the generation of fluent and informative text. In contrast, isolated or weakly connected nodes often represent peripheral information, contributing less to the overall meaning. While this alignment is not quantified in this section, it is consistently observed across the sampled examples, providing observational support that structural completeness often correlates with richer semantic coverage in the graph-to-text generation setting.

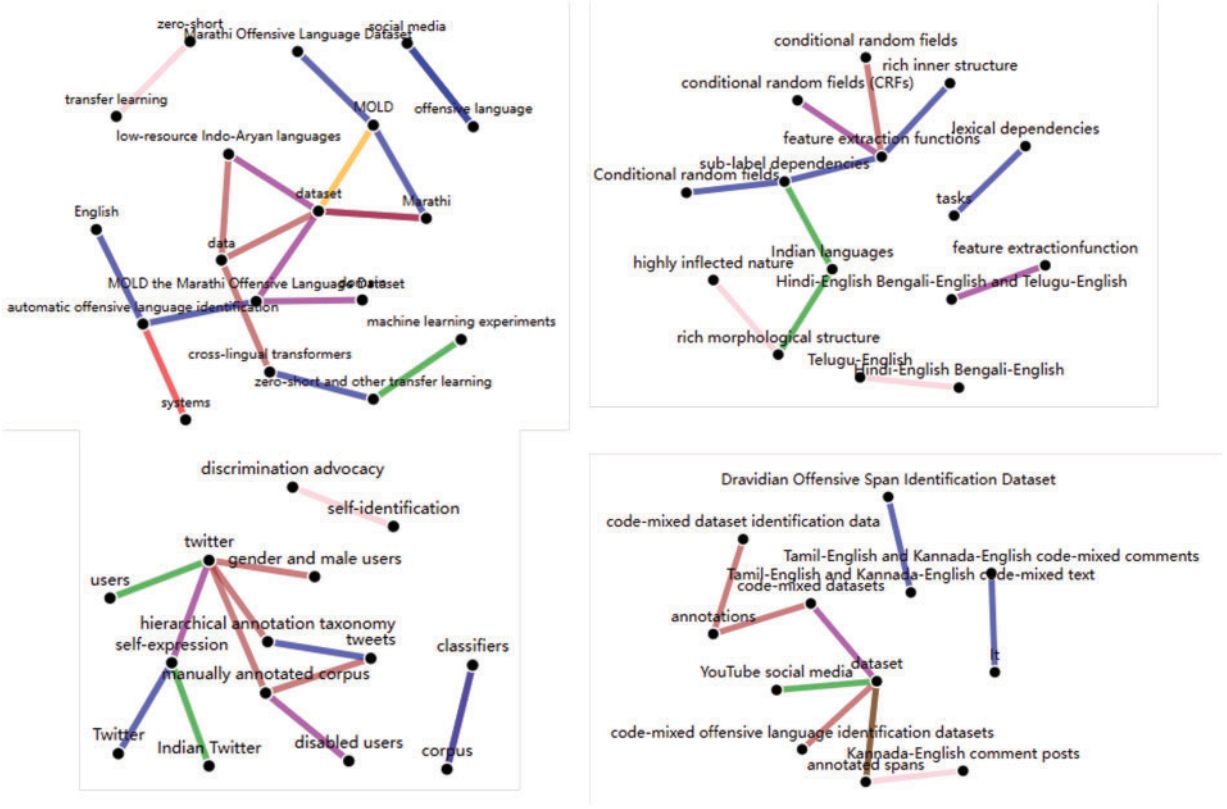

Figure 8: Four randomly sampled knowledge graphs from the top-300 ranked set

4.8 Human Evaluation

To verify that our proposed ranking method truly reflects human preferences, we conducted a focused human study. We first partitioned the automatically ranked list into the top-300 and bottom-300. From each partition, we randomly sampled ten graphs, resulting in a balanced set of 20 KG-text pairs. Each pair was scored on a 1–5 Likert scale based on two questions: **Q1 Graph Quality:** "Is the KG structurally complete and error-free?" **Q2 Text Fidelity:** "Does the generated text faithfully describe the knowledge graph's contents?" For each question, we averaged the annotators' scores per item, computed the group means, and applied an independent two-sample t-test between the top and bottom groups.

Table 6 shows that graphs in the top-ranked tranche receive markedly higher human judgments than those in the bottom tranche for both structural soundness and textual faithfulness. The differences are statistically significant $p < 0.0001$ in both structural quality and textual fidelity. These results confirm that our automatic dual-perspective framework aligns closely with human perceptions of knowledge-graph quality and generation fidelity, providing empirical support for its reliability in downstream selection scenarios.

Table 6: Human evaluation of sampled top and bottom ranked KG-text pairs

Metric	Top-300 sample (n = 10)	Bottom-300 sample (n = 10)	*t*-Value	*p*-Value
Graph quality	4.32 ± 0.21	2.11 ± 0.18	7.99	$p < 0.0001$
Text fidelity	4.07 ± 0.26	1.95 ± 0.13	7.29	$p < 0.0001$

5 Conclusion

High-quality knowledge graphs underpin effective and reliable graph-to-text generation. Despite this, research in this area remains scarce. Earlier methods leaned heavily on broad quality measures, often missing what generation tasks demand. This study addresses this gap by pointing out that existing methods frequently overlook structural and semantic characteristics essential for producing high-quality text. We propose an evaluation framework for graph-to-text generation, which assesses knowledge graphs from structural and semantic perspectives. We introduce the WIEM for structural evaluation, which quantifies completeness based on the agreement between an LLM and a relation extraction model. We convert KGs into natural language for semantic evaluation and measure the similarity between the generated and reference texts to assess whether the graph preserves the intended meaning. These two modules are jointly applied using a Top-K union method, allowing us to identify knowledge graphs best suited for generation. Experiments on a scientific abstract dataset and the general-domain knowledge graph dataset ACE05 demonstrate that incorporating the selected KGs as support examples in in-context learning consistently enhances generation quality. Further evaluations, including random edge perturbation tests, confirm the robustness and consistency of our framework under different conditions.

Although it works well, the framework depends on models and data built for the scientific domain. Applying it to other domains may require adjustments to the models. Future work will explore broader domain adaptation and investigate whether the approach remains effective across different text generation models. Additionally, we will investigate more constrained decoding strategies, such as constrained beam search or entity-aware decoding, to further suppress external hallucinations and enhance the semantic evaluation's fidelity to the input knowledge graph.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Writing—review & editing, Conceptualization of this study, Methodology, Experimentation, Formal analysis, Software: Haotong Wang; Review & editing, Methodology, Experimentation, Formal analysis: Liyan Wang; Research topic, Review & editing, Methodology, Formal analysis, Resources, Supervision: Yves Lepage. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available on the website: http://lepage-lab.ips.waseda.ac.jp/projects/scientific-writing-aid (accessed on 28 April 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst. 1996 Mar;12(4):5–33. doi:10.1080/07421222.1996.11518099.
2. Ji S, Pan S, Cambria E, Marttinen P, Yu PS. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst. 2022;33(2):494–514. doi:10.1109/TNNLS.2021.3070843.
3. Schneider P, Schopf T, Vladika J, Galkin M, Simperl E, Matthes F. A decade of knowledge graphs in natural language processing: a survey. In: He Y, Ji H, Li S, Liu Y, Chang CH, editors. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2022; Online: Association for Computational Linguistics. p. 601–14.
4. Xue B, Zou L. Knowledge graph quality management: a comprehensive survey. IEEE Trans Knowl Data Eng. 2023;35(5):4969–88.
5. Yu W, Zhu C, Li Z, Hu Z, Wang Q, Ji H, et al. A survey of knowledge-enhanced text generation. ACM Comput Surv. 2022;54(11):227. doi:10.1145/3512467.
6. Dong C, Li Y, Gong H, Chen M, Li J, Shen Y, et al. A survey of natural language generation. ACM Comput Surv. 2022 Dec;55(8):173. doi:10.1145/3554727.
7. Ribeiro LFR, Schmitt M, Schütze H, Gurevych I. Investigating pretrained language models for graph-to-text generation. In: Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI; 2021; Online: Association for Computational Linguistics. p. 211–27.
8. Colas A, Alvandipour M, Wang DZ. GAP: a graph-aware language model framework for knowledge graph-to-text generation. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022; Gyeongju, Republic of Korea: International Committee on Computational Linguistics. p. 5755–69.
9. Koncel-Kedziorski R, Bekal D, Luan Y, Lapata M, Hajishirzi H. Text generation from knowledge graphs with graph transformers. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019; Minneapolis, Minnesota: Association for Computational Linguistics. p. 2284–93.
10. Li Z, Huang W, Gong X, Luo X, Xiao K, Deng H, et al. Decoupled semantic graph neural network for knowledge graph embedding. Neurocomputing. 2025;611(5):128614. doi:10.1016/j.neucom.2024.128614.
11. Morrison D, Bedinger M, Beevers L, McClymont K. Exploring the raison d'etre behind metric selection in network analysis: a systematic review. Appl Netw Sci. 2022;7(1):50. doi:10.1007/s41109-022-00476-w.
12. Issa S, Adekunle O, Hamdi F, Cherfi SSS, Dumontier M, Zaveri A. Knowledge graph completeness: a systematic literature review. IEEE Access. 2021;9:31322–39. doi:10.1109/ACCESS.2021.3056622.
13. Li Z, Chen L, Jian Y, Wang H, Zhao Y, Zhang M, et al. Aggregation or separation? Adaptive embedding message passing for knowledge graph completion. Inf Sci. 2025;691(11):121639. doi:10.1016/j.ins.2024.121639.
14. Axelsson A, Skantze G. Using large language models for zero-shot natural language generation from knowledge graphs. In: Gatt A, Gardent C, Cripwell L, Belz A, Borg C, Erdem A et al., editors. Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023); 2023; Prague, Czech Republic: Association for Computational Linguistics. p. 39–54.
15. Zhong Z, Chen D. A frustratingly easy approach for entity and relation extraction. In: Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tur D, Beltagy I, Bethard S, et al. editors. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021; Online: Association for Computational Linguistics. p. 50–61.
16. Dong Q, Li L, Dai D, Zheng C, Ma J, Li R, et al. A survey on in-context learning. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; 2024; Miami, FL, USA: Association for Computational Linguistics. p. 1107–28.

17. Ibrahim N, Aboulela S, Ibrahim A, Kashef R. A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges. Discover Artif Intell. 2024;4(1):76. doi:10.1007/s44163-024-00175-8.
18. Wang J, Liu YJ, Li P, Lin Z, Sindakis S, Aggarwal S. Overview of Data quality: Examining the dimensions, antecedents, and impacts of data quality. J Knowl Econ. 2023;15(1):1159–78. doi:10.1007/s13132-022-01096-6.
19. Saxena A, Iyengar S. Centrality measures in complex networks: a survey. arXiv:2011.07190. 2020.
20. Cimiano P, Paulheim H. Knowledge graph refinement: a survey of approaches and evaluation methods. SemantWeb. 2017 Jan;8(3):489–508.
21. Al Musawi AF, Roy S, Ghosh P. Examining indicators of complex network vulnerability across diverse attack scenarios. Sci Rep. 2023;13(1):18208. doi:10.1038/s41598-023-45218-9.
22. Laita A, Kotiaho JS, Monkkonen M. Graph-theoretic connectivity measures: what do they tell us about connectivity? Landsc Ecol. 2011;26:951–67.
23. Chen H, Cao G, Chen J, Ding J. A practical framework for evaluating the quality of knowledge graph. In: Zhu X, Qin B, Zhu X, Liu M, Qian L, editors. Knowledge graph and semantic computing: knowledge computing and language understanding. Singapore: Springer Singapore; 2019. p. 111–22.
24. Fanourakis N, Efthymiou V, Christophides V, Kotzinos D, Pitoura E, Stefanidis K. Structural bias in knowledge graphs for the entity alignment task. In: Pesquita C, Jiménez-Ruiz E, McCusker JP, Faria D, Dragoni M, Dimou A, et al. editors. ESWC. Vol. 13870 of lecture notes in computer science. Heidelberg, Germany: Springer; 2023. p. 72–90.
25. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020; 2020 Apr 26–30; Addis Ababa, Ethiopia; 2020.
26. Sellam T, Das D, Parikh A. BLEURT: learning robust metrics for text generation. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020; Online: Association for Computational Linguistics. p. 7881–92.
27. Luan Y, He L, Ostendorf M, Hajishirzi H. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, editors. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018; Brussels, Belgium: Association for Computational Linguistics. p. 3219–32.
28. Wang H, Lepage Y. Extraction-Augmented generation of scientific abstracts using knowledge graphs. IEEE Access. 2025;13:48775–91.
29. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The Llama 3 herd of models. arXiv: 2407.21783. 2024.
30. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74–81.