

Ontology-Free General-Domain Knowledge Graph-to-Text Generation Dataset Synthesis using Large Language Model

Daehui Kim^{1,2}, Deokhyung Kang², Sangwon Ryu² and Gary Geunbae Lee^{2,3}

¹Agentic AI Lab, KT

²Graduate School of Artificial Intelligence, POSTECH

³Department of Computer Science and Engineering, POSTECH
{andrea0119, deokhk, ryusangwon, gblee}@postech.ac.kr

Abstract

Knowledge Graph-to-Text (G2T) generation involves verbalizing structured knowledge graphs into natural language text. Recent advancements in Pretrained Language Models (PLMs) have improved G2T performance, but their effectiveness depends on datasets with precise graph-text alignment. However, the scarcity of high-quality, general-domain G2T generation datasets restricts progress in the general-domain G2T generation research. To address this issue, we introduce **Wikipedia Ontology-Free Graph-text dataset (WikiOFGraph)**, a new large-scale G2T dataset generated using a novel method that leverages Large Language Model (LLM) and Data-QuestEval. Our new dataset, which contains 5.85M general-domain graph-text pairs, offers high graph-text consistency without relying on external ontologies. Experimental results demonstrate that PLM fine-tuned on WikiOFGraph outperforms those trained on other datasets across various evaluation metrics. Our method proves to be a scalable and effective solution for generating high-quality G2T data, significantly advancing the field of G2T generation.¹

1 Introduction

Knowledge Graph-to-Text Generation (G2T) is a task aimed at verbalizing knowledge graphs represented as a set of triplets in the form of (*subject, predicate, object*) into natural language text (Gatt and Kraemer, 2018; Lin et al., 2024). Recent advancements in Pretrained Language Models (PLMs), such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) have led to significant improvement in G2T when fine-tuned on G2T datasets (Kale and Rastogi, 2020).

However, fine-tuning PLMs requires a substantial amount of well-aligned G2T data (Kasner et al., 2023). Constructing G2T datasets is labor-intensive and expensive because understanding

¹Our code and data are available at: <https://github.com/daehuikim/WikiOFGraph>

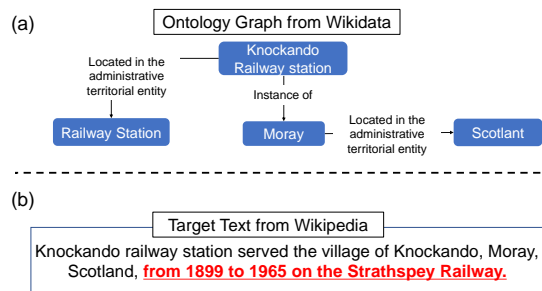


Figure 1: An example of a Graph-Text pair from existing ontology-based datasets. Although (a) and (b) are paired, the graph in (a) does not contain the information of the underlined text in (b), illustrating a common misalignment problem.

structured graph representations for various natural language sentences is challenging for human annotators. Therefore, previous research has mainly focused on small-scale domain-specific datasets (Gardent et al., 2017; Koncel-Kedziorski et al., 2019; Castro Ferreira et al., 2020; Nan et al., 2021).

To address this issue, researchers have attempted to automatically align Wikipedia texts with their corresponding graphs from ontologies to construct large-scale, general-domain G2T datasets (Elsahar et al., 2018; Jin et al., 2020; Agarwal et al., 2021; Wang et al., 2021; Mousavi et al., 2024). However, ontology-based datasets often suffer from graph-text misalignment, making it challenging to generate text accurately. For example, in Figure 1 (b), the phrase "from 1899 to 1965 on the Strathspey railway" cannot be inferred solely from the graph shown in Figure 1 (a). Such discrepancy can be a major cause of the decline in G2T performance (Mousavi et al., 2024).

Recent advances in Large Language Models (LLMs) offer promising potential to address these challenges. Many studies have successfully leveraged LLMs to synthesize high-quality data for var-

ious tasks (Long et al., 2024). Several studies have attempted to generate graph-text paired data using LLMs (Josifoski et al., 2023; Han et al., 2024; Chen et al., 2024), but these attempts have yet to be thoroughly explored for the G2T generation task.

To address this issue, we introduce an effective method for generating high-quality G2T dataset that integrates LLM with Data-QuestEval (Rebuffel et al., 2021). We select three examples from the human-crafted dataset to facilitate In-Context Learning (ICL) for the LLM, enabling it to extract graph representations from the given sentences. To ensure high consistency between the extracted graph representations and text, we utilize Data-QuestEval for data curation. Through this method, we create a new dataset called **Wikipedia Ontology-Free Graph-Text dataset (WikiOFGraph)**, a 5.85M G2T data that covers a broad range of Wikipedia articles without relying on ontologies.

Through comprehensive analyses, we demonstrate that our dataset achieves graph-text consistency comparable to that of a fully human-crafted dataset while significantly surpassing other ontology-based general-domain G2T datasets. Experimental results demonstrate that a PLM fine-tuned on WikiOFGraph outperforms those trained on existing datasets across the human expert-crafted GenWiki (Jin et al., 2020) test set and LLM-synthesized WikiOFGraph test set. This highlights the suitability of our dataset for building G2T systems that perform well across general domains, making it more effective than other datasets. We further demonstrate the effectiveness of Data-QuestEval filtering through additional experiments. In summary, our key contributions are as follows:

- (i) We introduce an effective method for synthesizing high-quality G2T dataset. Our approach is independent of proprietary LLMs or ontologies, making it reproducible and easily adaptable for various domains.
- (ii) We release a new G2T generation dataset called **WikiOFGraph**. Our comprehensive analyses indicate that it offers high graph-text consistency, comparable to a human-crafted dataset, with 5.85M samples covering the entire spectrum of Wikipedia.
- (iii) We demonstrate the effectiveness of Data-QuestEval filtering through additional experiments and case study.

2 Related Works

Graph-to-Text generation Traditionally, researchers tackled G2T generation using templates designed to expressing predicates into pre-defined statements (Wiseman et al., 2018; Kasner and Dusek, 2022; Xiang et al., 2022; Vejvar and Fujimoto, 2023). Template-based approaches have low hallucination rates but are labor-intensive in creating various templates, and they often struggle with producing fluent sentences from complex graphs (Kasner and Dušek, 2020).

To address these limitations, researchers have leveraged neural encoder-decoder architectures (Wiseman et al., 2017; Beck et al., 2018; Nie et al., 2018; Puduppully et al., 2019; Iso et al., 2019) which convert graph representations into vector embeddings interpretable by neural models. Unlike previous approaches, these approaches leverage an end-to-end G2T generation paradigm that does not require pre-defined templates.

The introduction of transformer (Vaswani et al., 2017)-based PLMs has significantly advanced G2T generation, leading to much better performance than earlier methods. Kale and Rastogi demonstrated substantial improvements by fine-tuning these PLMs on G2T generation tasks. This strategy has since been adopted in subsequent studies (Ribeiro et al., 2021; Jolly et al., 2022; Mehta et al., 2022; Han and Shareghi, 2022). However, this new paradigm requires substantial amounts of well-aligned G2T data.

G2T generation datasets WebNLG (Gardent et al., 2017; Castro Ferreira et al., 2020) is a fully human-crafted dataset based on the DBpedia (Mendes et al., 2012), making it a representative benchmark for G2T generation tasks due to its precise graph-text alignment. However, creating human-crafted datasets is resource-intensive and limits scale. These limitations on scale and diversity have led researchers to develop large-scale G2T datasets automatically.

GenWiki (Jin et al., 2020) provides a large-scale, general-domain G2T dataset with 1.3 M samples. It includes fine and full splits, determined by the F1-score-based alignment between text and graph entities, and 1,000 human-crafted test samples. TekGen (Agarwal et al., 2021) combines Wikipedia and Wikidata (Vrandečić and Krötzsch, 2014) to create a large-scale, general-domain G2T dataset using a distant supervision strategy. This approach allowed for the creation of large datasets but led to insufficient alignment between graph representa-

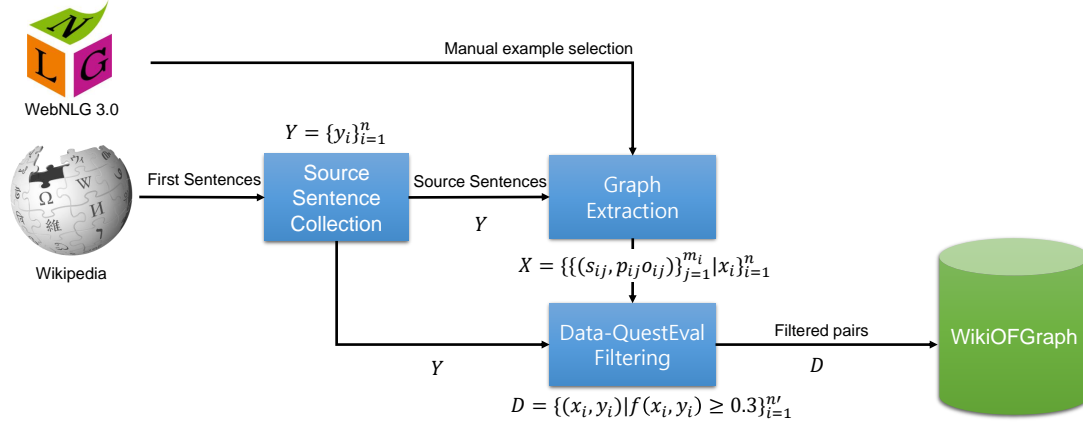


Figure 2: Method for constructing the **WikiOFGraph**. Source sentences are collected from Wikipedia. Graph representations are then extracted using an LLM through in-context learning, guided by manually selected examples from the WebNLG. Data-QuestEval Filtering curates graph-text pairs compiled into the **WikiOFGraph**.

tions and target texts. LAGRANGE (Mousavi et al., 2024), a 3.0M data, aligns Wikidata and Wikipedia data and incorporates second-hop reasoning and triplet augmentation.

However, these attempts rely on text aligning algorithms based on ontologies with insufficient information, leading to inadequate graph-text alignment. In contrast, our method generates graphs directly from the source text, better capturing the text’s content. Our method is simple and scalable, enabling the creation of significantly larger datasets with greater domain diversity than existing ontology-based datasets.

LLM-driven data generation The natural language processing community has recently shifted its focus toward synthesizing high-quality, large-scale data using LLMs (Long et al., 2024). While LLM-generated data is easily applicable to discriminative tasks (Whitehouse et al., 2023; Ghosh et al., 2024), its effectiveness in generative tasks diminishes when the distribution of the generated data differs from existing data (Ding et al., 2024).

While several studies have explored using LLMs for generating graph-text pairs, their primary focus has been on tasks such as Text-to-Graph (T2G) generation rather than on G2T generation (Josifoski et al., 2023; Han et al., 2024; Mousavi et al., 2024; Chen et al., 2024). In contrast, our research focuses on the LLM-driven synthetic graph generation for creating G2T generation dataset. Moreover, our approach leverages Data-QuestEval for data filtering, ensuring high graph-text consistency without manual or heuristic adjustments.

3 WikiOFGraph

In this section, we introduce a method for creating **WikiOFGraph**. We first define the requirements necessary to address the limitations of existing G2T datasets (§3.1). We then outline the rule-based algorithm for gathering source sentences from Wikipedia (§3.2), as well as the process of extracting graphs through LLM (§3.3). Additionally, we describe our approach to utilizing Data-QuestEval (Rebuffel et al., 2021) for curating well-aligned graph-text pairs (§3.4).

3.1 Requirements

Graph-text consistency Good G2T data guarantees that the target text includes *“all”* and *“only”* information from the graph. If *“all”* information of the graph is not reflected in the target text, it can cause data omission problems for the model. If the target text does not includes *“only”* information from the graph, it can cause hallucination problems for the model. Therefore, graph-text consistency is essential to prevent omission and hallucination problems in the G2T generation task.

Domain diversity PLMs fine-tuned on datasets limited to a specific domain often struggle to handle samples from unseen domains (Keymanesh et al., 2022). This issue becomes apparent when PLMs are tested on new domains, where performance falls short compared to familiar domains (Castro Ferreira et al., 2020). Therefore, including a wide range of domains is crucial to developing a domain-adaptive G2T generation system.

Large scale The demand for substantial structured datasets is growing with the emergence of

language models trained on extremely large-scale structured data (Zhuang et al., 2024; Li et al., 2024). High-quality structured data, such as knowledge graph representations, is essential for training LLMs for structured data processing. However, creating such data is far more challenging than compiling unstructured corpora. Consequently, providing a large-scale G2T dataset could be key to advancing LLMs in handling structured data effectively.

3.2 Source Sentences Collection

To meet the requirements (§3.1), we select Wikipedia as the source of sentences. Wikipedia includes sentences with extensive factual knowledge from diverse domains. We employ a rule-based algorithm to extract the first sentence from every article on English Wikipedia because the first sentence of a Wikipedia article often encapsulates vital factual information. We apply several constraints to enhance the clarity of the factual context in the source sentences. We limit the length of sentences to 10-500 characters, exclude sentences starting with pronouns to avoid ambiguous entity extraction (e.g., 'it,' 'that'), and remove parenthetical explanations to prevent redundant expressions. Through this process, we collect 6.06M source sentences, denoted as $Y = \{y_i\}_{i=1}^n$ in Figure 2, where n represents the total number of source sentences.

3.3 Graph Extraction

We utilize the "Llama-3-70b-instruct-awq"² (Dubey et al., 2024) to extract graph representations directly from Y . We manually select three examples from the WebNLG (Castro Ferreira et al., 2020) dataset for in-context examples³. We then obtain a set of graph representations for each sentence, denoted as $X = \left\{ \{(s_{ij}, p_{ij}, o_{ij})\}_{j=1}^{m_i} \mid x_i \right\}_{i=1}^n$ in Figure 2. Here, n represents the total number of sentences, m_i denotes the number of triplets associated with the i -th sentence, and (s_{ij}, p_{ij}, o_{ij}) are the subject, predicate, and object of the j -th triplet in the i -th sentence, respectively. Through this process, we obtain 6.06M pairs of source sentences and their corresponding generated graph representations.

3.4 Data-QuestEval Filtering

Data-QuestEval (Rebuffel et al., 2021) is a framework that measures the accuracy of predicted texts

²Available at: [casperhansen/llama-3-70b-instruct-awq](https://casperhansen.github.io/llama-3-70b-instruct-awq/)

³For implementation details, see the Appendix A

in Data-to-Text generation tasks by combining question generation and question-answering techniques to evaluate whether the *predicted sentence* accurately reflects the content of the *source data*. Data-QuestEval supports referenceless evaluation by relying solely on the *source data* and the *predicted sentence*. We adapt referenceless evaluation by using the X as the *source data* and the Y as the *predicted sentence*.

We apply the referenceless Data-QuestEval scoring $f(x_i, y_i)$ to evaluate the consistency between the generated graph representations and their corresponding source sentences. Only pairs with a score of $f(x_i, y_i) \geq 0.3$ are retained for further curation, as denoted by $D = \{(x_i, y_i) \mid f(x_i, y_i) \geq 0.3\}_{i=1}^{n'}$ in Figure 2, where n' represents the total number of curated graph-text pairs. Through this process, we obtain 5.95M⁴ samples, with less than 2% of the samples being filtered out. This result implies that a robust LLM can reliably extract graph representations containing "all" and "only" information from given texts.

4 Data Analysis

We compare WikiOFGraph with existing datasets. We explore the scale and domain diversity of the data through quantitative analysis (§4.1). We then assess the graph-text consistency using human evaluators and GPT-4o (OpenAI et al., 2024) (§4.2).

We utilize four representative G2T datasets for our comparative analysis: human-crafted WebNLG (Castro Ferreira et al., 2020) and three automatically generated ontology-based datasets—GenWiki⁵ (Jin et al., 2020), TekGen (Agarwal et al., 2021), and LAGRANGE (Mousavi et al., 2024). Details of the datasets are provided in the appendix (§Appendix D).

4.1 Quantitative Analysis

Dataset	# samples	# unique predicate	# unique entity	# triplet (m/M/avg)
WebNLG	35K	372	3.2K	1/7/2.96
GenWiki	680K	287	86.6K	1/10/2.64
TekGen	6.31M	50,861	4.3M	1/54/1.73
LAGRANGE	3.07M	1,167	2.9M	1/135/4.02
WikiOFGraph (Ours)	5.85M	140,733	8.2M	1/17/3.62

Table 1: Training set statistics for comparative analysis. # triplet (m/M/avg) indicates the minimum, maximum, and average number of triplets per sample.

Domain diversity & dataset volume Table 1 provides a detailed comparison of statistics across

⁴We employ 5.85M samples for the training split and 100K samples for the test split.

⁵We employ the GenWiki_{FINE} training split

various G2T datasets. WikiOFGraph includes a remarkably higher number of unique predicates and unique entities than other datasets. This diversity in predicates and entities is anticipated to enhance domain generalization in G2T generation tasks. In addition, WikiOFGraph is advantageous in terms of scale, containing a substantial 5.85 M data pairs, which makes it a valuable resource for future research in G2T generation tasks.

In contrast, WebNLG contains the smallest number of unique predicates, entities, and samples. This limitation is primarily due to its entirely human-crafted nature, which makes it challenging to produce a large volume of data.

GenWiki and LAGRANGE also have fewer unique predicates and entities. This is because the underlying ontology graphs in these datasets cover a relatively narrow scope. Also, these datasets are designed to support G2T and T2G generation tasks, limiting the variety of available predicate expressions.

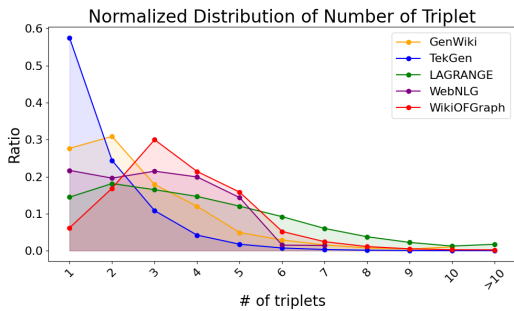


Figure 3: Normalized distribution of the number of triplets in each dataset.

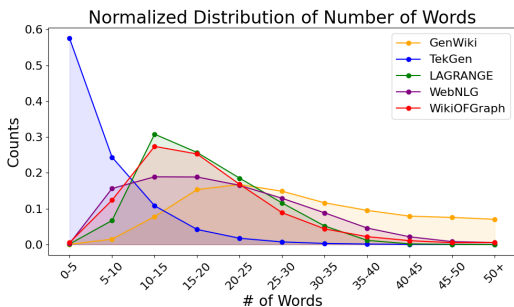


Figure 4: Normalized distribution of the number of words in each dataset.

Statistical comparison of dataset compositions

To provide more detailed comparisons, we conduct a statistical analysis of these datasets. Figure 3 shows the distribution of samples categorized by the number of triplets, while Figure 4 presents the distribution of samples categorized by word

count in increments of five. WikiOFGraph demonstrates a balanced and consistent dataset structure, with the distribution of triplets and word counts closely aligning. The similarity in the shapes of these two distributions, which differs from those of other datasets, suggests a consistent alignment between the graph and the target text.

Despite TekGen showing good alignment between triplet and word count distributions, most of its samples consist of two or fewer triplets or contain fewer than ten words, indicating that most samples are relatively short. GenWiki predominantly consists of samples with three or fewer triplets; however, the relatively high proportion of samples with more than 20 words suggests a surface-level inconsistency between the graph representations and the corresponding texts.

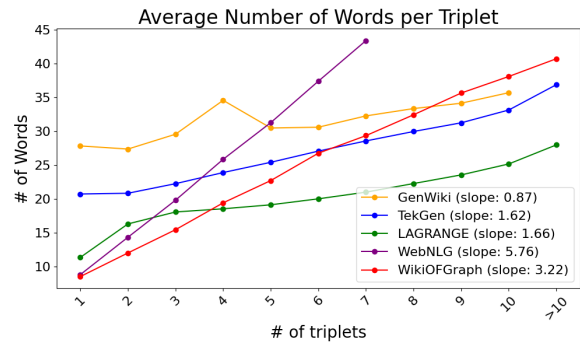


Figure 5: Average number of words per number of triplets across different datasets.

Average words per triplet We also visualize the average number of words per sample for each dataset, categorized by the number of triplets, as shown in Figure 5. WikiOFGraph exhibits an average word increase rate of 3.22 per triplet, which closely aligns with the average of 3.62 triplets per sample, as reported in Table 1. This suggests that WikiOFGraph provides consistent information with minimal bias relative to the number of triplets.

WebNLG exhibits a relatively high word increase rate of 5.76 per triplet compared to other datasets. However, this rate contrasts with the average of 2.96 triplets per sample, as reported in Table 1. This discrepancy arises because most WebNLG samples consist of short sentences and few triplets.

On the other hand, the automatically generated ontology-based datasets show much lower word increase rates per triplet, ranging from 0.87 to 1.66. These lower rates suggest that each triplet in these datasets is associated with fewer words, leading to

less detailed representations per triplet.

4.2 Qualitative Analysis

To further analyze graph-text consistency, we conduct a qualitative analysis employing human and GPT-4o. We provide 30 samples per dataset, each evaluated by five human reviewers who assess the same set of samples, and an additional 1,000 samples per dataset are evaluated by GPT-4o. Since some datasets have a high ratio of samples with a low number of triplets, we categorize the samples by the number of triplets and randomly sample them within each range to ensure a diverse evaluation across different sample lengths.

Our focus is to assess whether the information of graph representations (triplets) are accurately reflected in the target text. To achieve this, we ask evaluators to review graph-text pairs and identify 1) unused triplets that are not used to generate the text and 2) parts of the text that could not be guessed from the triplets.

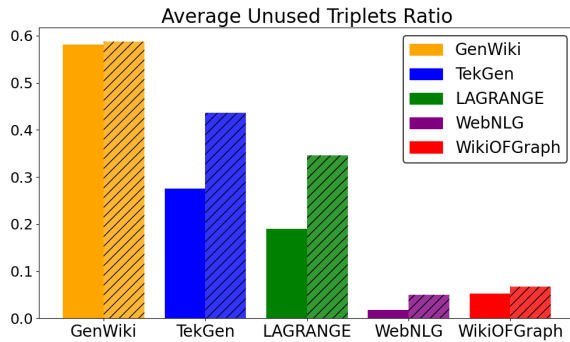


Figure 6: Number of unused triplet ratio

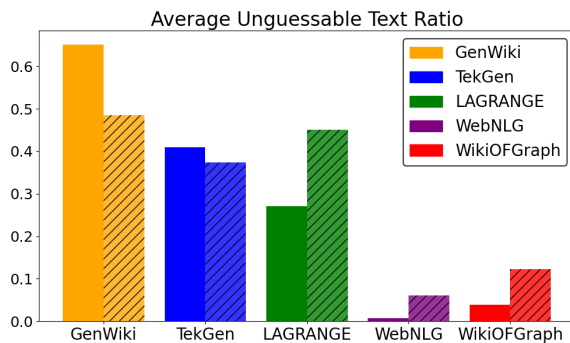


Figure 7: Length of unguessable text ratio

Graph-text consistency Figure 6 presents the average ratio of unused triplets across different datasets. The plain bars represent the evaluations conducted by human evaluators, while the hatched bars indicate the assessment results from GPT-4o.

When comparing automatically generated datasets, the average ratio of unused triplets is highest in GenWiki, followed by TekGen, LAGRANGE, and WikiOFGraph. Notably, WikiOFGraph is comparable to WebNLG, a fully human-crafted dataset. This observation suggests that WikiOFGraph incorporates triplet information into the text better than automatically generated ontology-based datasets.

Figure 7 presents the average ratio of unguessable text across different datasets. Similarly, GenWiki, TekGen, and LAGRANGE exhibit relatively high amounts of unguessable text, indicating a significant presence of text that cannot be generated from the graph information alone. In contrast, WikiOFGraph contains a low amount of unguessable text, even comparable to WebNLG. This observation suggests that WikiOFGraph has significantly higher graph-text consistency than automatically generated ontology-based G2T datasets. Details regarding the evaluation process can be found in the appendix (§Appendix B).

5 Experiments

We provide a comprehensive overview of our experimental setup and results. We first outline our experimental settings (§5.1). We then describe the evaluation metrics used to assess the model’s G2T generation performance (§5.2). We describe the evaluation data selection, focusing on ensuring a broad topic range and high reliability (§5.3).

5.1 Experimental Settings

We choose the T5-large (Raffel et al., 2020), which has 770M parameters, for our backbone model. The backbone model is trained to take *source triplets* as input and predict a *target text*. We use special tokens such as ‘<S>’, ‘<P>’, and ‘<O>’ to distinguish triplet elements. For example, a *source triplets* like “(<S> Arròs negre| <P> country| <O> Spain), (<S> Spain| <P> ethnic Group| <O> Spaniards)” is used to predict a *target text* such as “Arros negre is from Spain where Spaniards are an ethnic group.” We utilize Huggingface’s transformers (Wolf et al., 2020) library for our experiments. We use the cross-entropy loss function during training. Implementation details and hyperparameter settings are provided in the appendix (§Appendix C).

5.2 Evaluation Metrics

We employ representative metrics commonly used to evaluate the performance of G2T generation tasks. We utilize several metrics, including BLEU

	BLEU (↑)	ChrF++ (↑)	METEOR (↑)	TER (↓)	BLEURT (↑)	ROUGE-L (↑)	BertScore-F1 (↑)
WebNLG	35.97	65.76	39.69	49.71	0.44	64.41	95.56
GenWiki	36.14	56.96	33.80	89.77	0.13	58.13	92.85
TekGen	39.25	63.43	38.19	75.15	0.30	64.16	94.13
LAGRANGE	40.40	67.51	40.70	47.15	0.43	65.26	95.71
WikiOFGraph (Ours)	45.85	69.61	42.17	43.38	0.46	70.10	95.89

Table 2: Evaluation result on GenWiki test set.

	BLEU (↑)	ChrF++ (↑)	METEOR (↑)	TER (↓)	BLEURT (↑)	ROUGE-L (↑)	BertScore-F1 (↑)
WebNLG	21.91	35.97	39.69	49.71	0.44	62.88	93.81
GenWiki	23.42	48.30	33.17	74.89	0.14	61.92	92.14
TekGen	38.45	62.28	41.39	55.01	0.45	73.08	94.62
LAGRANGE	39.37	63.33	42.85	46.04	0.51	74.68	95.08
WikiOFGraph (Ours)	69.27	82.63	51.59	17.44	0.71	86.46	98.29

Table 3: Evaluation result on WikiOFGraph test set.

(Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ChrF++ (Popović, 2015), TER (Snover et al., 2006), BLEURT (Sellam et al., 2020) and ROUGE-L (Lin, 2004), specifically to evaluate the fluency of the predicted sentences. To assess semantic accuracy, we employ BERTScore-F1 (Zhang et al., 2020).

5.3 Evaluation Data Selection

We utilize evaluation datasets that cover a broad range of topics while maintaining high reliability to assess the model’s general domain G2T capabilities. First, we choose the GenWiki (Jin et al., 2020) test set, consisting of 1,000 samples generated by human experts, for its high level of trustworthiness. Next, we select 100,000 samples from the WikiOFGraph dataset not included in the training split as an additional evaluation dataset. Although WikiOFGraph consists of synthesized data, our earlier analyses indicate that it possesses high domain diversity and strong graph-text consistency. This makes it an appropriate choice for evaluating the model’s G2T generation capabilities in the general domain.

6 Results and Discussions

Table 2 and 3 present the performances of the T5-large model fine-tuned on five different datasets, with the best values highlighted in bold. The results demonstrate that fine-tuning with WikiOFGraph consistently achieves the highest performance across various metrics on both test sets.

Training on domain-specific, human-crafted datasets like WebNLG leads to lower performance in general domain G2T generation tasks due to the limited size and narrow coverage of the training

data. On the other hand, training on datasets that cover a broader range of domains but are automatically generated based on ontology also leads to lower performance due to reduced graph-to-text consistency. This highlights the importance of dataset scale, domain diversity, and graph-text consistency in achieving robust performance in general domain G2T generation as mentioned in §3.1.

Table 3 highlights a more apparent performance gap between the T5-large model fine-tuned on WikiOFGraph and those fine-tuned on other datasets. This performance gap is anticipated due to the similar distribution of test samples to WikiOFGraph, yet the difference is considerably significant. These results demonstrate that WikiOFGraph is significantly more effective than other G2T datasets for general domain G2T generation.

7 Effectiveness of Data-QuestEval

Due to the extremely low ratio of filtered samples (less than 2%), observing significant performance differences from Data-QuestEval curation is challenging. To further investigate the impact of Data-QuestEval, we conduct additional experiments with controlled sample sizes (§7.1). We then conduct a case study to verify whether Data-QuestEval effectively excludes misaligned graph-text pairs (§7.2).

7.1 Impact of Filtered Sample Ratios

Curated:Noise	BLEU (↑)	METEOR (↑)	Rouge-L (↑)	BertScore-F1 (↑)
0:100	43.21	40.62	66.11	95.56
25:75	44.25	41.01	67.47	95.79
50:50	44.57	41.22	67.88	95.77
75:25	44.75	41.16	68.38	95.83
100:0	45.00	41.37	69.32	96.00

Table 5: Impact of *Curated:Noise* sample ratios on model performance.

Status	Graph Representations	Text	Data-QuestEval Score
Incomplete Text Error	(<S>Piano sonata no.1 <P>TypeI <O>Music composition)	The Piano Sonata No.	0.1123
Incomplete Text Error	(<S>Dübs crane tankI <P>Modell <O>No.)	Dübs crane tank No.	0.1097
Ambiguous Pronoun Error	(<S>Mnl-2I <P>RoundI <O>Qualification round)	This was the qualification round for MNL-2.	0.1270
Ambiguous Pronoun Error	(<S>SI <P>Namel <O>O) (<S>SI <P>OperatorI <O>O)	This is the name of an aerial lift, as well as its operator.	0.1142
Properly Generated	(<S>Domenico PucciniI <P>studied UnderI <O>Giovanni Paisiello)	Domenico Puccini studied for a time under Giovanni Paisiello.	0.4999
Properly Generated	(<S>Dennis HamiltonI <P>signed Datal <O>October 21, 1967) (<S>Dennis HamiltonI <P>signed ByI <O>Los Angeles Lakers)	October 21, 1967 The Los Angeles Lakers signed Dennis Hamilton as a free agent.	0.6768
Properly Generated	(<S>Double Hill StationI <P>locationI <O>up the Rakaia River)	Double Hill Station, located up the Rakaia River.	0.8929

Table 4: Examples of Graph-Text pairs with corresponding Data-QuestEval Scores, ranging from 0 to 1. Text highlighted in red indicates potential errors or areas of concern.

We define ‘curated’ samples as those scoring 0.3 or higher in the Data-QuestEval filtering, while samples scoring below 0.3 are classified as ‘noise’. We randomly select these ‘curated’ samples from WikiOFGraph, while the ‘noise’ samples are from the filtered-out samples. Given WikiOFGraph’s large scale, increasing the ratio of noise directly in the dataset is challenging. Therefore, we fix the train split at 50,000 samples and vary the ratio of ‘curated’ to ‘noise’ samples. The experimental settings used for these experiments follows the procedure outlined in §5.1.

As shown in Table 5, the evaluation result of the fine-tuned T5-large model improves proportionally as the ratio of curated samples increases. Notably, all key metrics—including BLEU, METEOR, Rouge-L, and BertScore-F1—exhibit consistent improvement, indicating a comprehensive enhancement in model performance. This result indicates that our approach to measuring graph-text consistency through Data-QuestEval filtering is effective, particularly in scenarios with limited access to high-quality training data.

7.2 Case Study

We conduct a case study to understand the low graph-text consistency problems in samples filtered-out by Data-QuestEval. Through the case study, we recognize that the filtered-out samples exhibit two representative types of problems. We list these error types with examples along with their corresponding Data-QuestEval scores in Table 4.

The first error type arises from incomplete source sentences. For instance, a sentence like “*Piano Sonata No. 1, the default title for a composer’s first (or only) piano sonata, may refer to:*” can be split as an incomplete sentence such as “*Piano Sonata No.*” due to its punctuation mark, which causes confusion in the *sentence splitter*. These incomplete sentences make it difficult to extract graph representations accurately.

The second error type occurs due to the use of ambiguous pronouns in the source sentence. Sentences containing demonstratives like “*This*” or other ambiguous pronouns make it challenging to determine the subject-object structure accurately.

Lastly, we observe that most well-structured sentences, like “*October 21, 1967 The Los Angeles Lakers signed Dennis Hamilton as a free agent.*”, generally result in consistent graph representations.

These results highlight the importance of using unambiguous sentences for graph extraction while demonstrating that Data-QuestEval filtering identifies ‘noise’ samples from the G2T dataset.

8 Conclusion

This study introduces WikiOFGraph, a large-scale G2T generation dataset with 5.85M samples covering the entire Wikipedia domain. To address the issue of graph-text misalignment commonly found in ontology-based datasets, we propose a novel method leveraging LLM and Data-QuestEval to generate high-quality graph-text pairs. Comprehensive analyses reveal that WikiOFGraph achieves **graph-text consistency** comparable to the fully human-crafted dataset, while also exhibiting **large scale** and extensive **domain diversity**. Comparisons with representative G2T datasets demonstrate that fine-tuning PLMs with WikiOFGraph significantly enhances their ability to perform general domain G2T tasks. Additional experiments and a case study demonstrate the effectiveness of Data-QuestEval in ensuring high-quality graph-text alignments, reinforcing its value in data curation. Our approach provides a scalable and efficient method for generating high-quality G2T data without relying on proprietary LLMs, external ontologies or extensive human involvement, making it reproducible for advancing G2T generation.

Acknowledgements

This work was supported by Smart HealthCare Program(www.kipot.or.kr) funded by the Korean National Police Agency(KNPA, Korea) [Project Name: Development of an Intelligent Big Data Integrated Platform for Police Officers' Personalized Healthcare / Project Number: 220222M01].

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-2020-0-01789) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

Limitations

Multilingual extension Although our method is easily reproducible across various domains using open-source LLMs and public source texts, the scope of our study focuses mainly on English. Future research could extend the study's multilingual capabilities by employing LLMs with multilingual capabilities and utilizing source texts from multilingual corpora.

Room for improvement We apply various empirical rules of thumb, particularly in areas such as prompt engineering and sampling strategies. While our approach significantly outperforms existing datasets in meeting key requirements, there remains room for improvement, particularly in refining prompts, optimizing sampling parameters for graph extraction, and selecting the most effective LLM. Additionally, since Data-QuestEval might not be the most optimal measure, exploring alternative methods to assess the consistency between the generated graph representations and the source text could be a valuable direction for future work.

Expanding directions Our work focuses primarily on the direction of G2T generation. Building a dataset that also supports T2G generation poses challenges, such as the need to select predicate expressions from a predefined vocabulary, which could limit flexibility. Future work could address this by reassigning predicate expressions within the WikiOFGGraph dataset, potentially creating a more adaptable framework for T2G tasks.

Data contamination concerns One important limitation not yet addressed in our study is the potential for data contamination. Since the LLMs

used in our experiments are pre-trained on publicly accessible datasets like Wikipedia, they may already be familiar with the data used in our graph extraction tasks. To thoroughly assess whether our method is robust against data contamination, future research should replicate our experiments on a large corpus that is not part of the LLM's pre-training data.

Ethical Considerations

We prioritize transparency and reproducibility in our research by using widely accessible resources. We utilize the open-source LLM *Llama3-70b-instruct-awq*⁶ and the publicly available Wikipedia⁷ dataset, adhering to their respective licenses.

For human evaluation, we conduct the assessments through Upwork⁸, ensuring participants receive appropriate compensation for their efforts. Each evaluator is paid a fixed rate of \$60 for their work, which involves evaluating five datasets for approximately three hours. This compensation is determined to be fair and in line with industry standards, reflecting our commitment to ethical research practices and the fair treatment of participants.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. [Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

⁶**WikiOFGGraph** is built with Meta Llama3; Meta Llama 3 is licensed under the Meta Llama 3 Community License, Copyright © Meta Platforms, Inc. All Rights Reserved. More details can be found at: [Llama3 License](#)

⁷Wikipedia is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA 3.0). The specific terms of this license can be found at: [Wikipedia License](#)

⁸<https://www.upwork.com>

- Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. [Graph-to-sequence learning using gated graph neural networks](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283, Melbourne, Australia. Association for Computational Linguistics.
- Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. [The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results \(WebNLG+ 2020\)](#). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.
- Hanzhu Chen, Xu Shen, Qitan Lv, Jie Wang, Xiaoqi Ni, and Jieping Ye. 2024. [SAC-KG: Exploiting large language models as skilled automatic constructors for domain knowledge graph](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4345–4360, Bangkok, Thailand. Association for Computational Linguistics.
- Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. [Data augmentation using large language models: Data perspectives, learning paradigms and challenges](#).
- Abhimanyu Dubey et al. 2024. [The llama 3 herd of models](#).
- Hady Elsahar, Pavlos Vougiouklis, Arslan Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. [T-REx: A large scale alignment of natural language with knowledge base triples](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. [The WebNLG challenge: Generating text from RDF data](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.
- Sreyan Ghosh, Utkarsh Tyagi, Sonal Kumar, Chandra Kiran Evuru, Ramaneswaran S, S Sakshi, and Dinesh Manocha. 2024. [ABEX: Data augmentation for low-resource NLU via expanding abstract descriptions](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 726–748, Bangkok, Thailand. Association for Computational Linguistics.
- Jiuzhou Han, Nigel Collier, Wray Buntine, and Ehsan Shareghi. 2024. [PiVe: Prompting with iterative verification improving graph-based generative capability of LLMs](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6702–6718, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Jiuzhou Han and Ehsan Shareghi. 2022. [Self-supervised graph masking pre-training for graph-to-text generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4845–4853, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Hayate Iso, Yui Uehara, Tatsuya Ishigaki, Hiroshi Noji, Eiji Aramaki, Ichiro Kobayashi, Yusuke Miyao, Naoaki Okazaki, and Hiroya Takamura. 2019. [Learning to select, track, and generate for data-to-text](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2102–2113, Florence, Italy. Association for Computational Linguistics.
- Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020. [GenWiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2398–2409, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Shailza Jolly, Zi Zhang, Andreas Dengel, and Lili Mou. 2022. [Search and learn: Improving semantic coverage for data-to-text generation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:10858–10866.
- Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. [Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.
- Mihir Kale and Abhinav Rastogi. 2020. [Text-to-text pre-training for data-to-text tasks](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.
- Zdeněk Kasner and Ondřej Dušek. 2020. [Data-to-text generation with iterative text editing](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 60–67, Dublin, Ireland. Association for Computational Linguistics.

- Zdeněk Kasner and Ondrej Dusek. 2022. [Neural pipeline for zero-shot data-to-text generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3914–3932, Dublin, Ireland. Association for Computational Linguistics.
- Zdeněk Kasner, Ioannis Konstas, and Ondrej Dusek. 2023. [Mind the labels: Describing relations in knowledge graphs with pretrained models](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2398–2415, Dubrovnik, Croatia. Association for Computational Linguistics.
- Moniba Keymanesh, Adrian Benton, and Mark Dredze. 2022. [What makes data-to-text generation hard for pretrained language models?](#) In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 539–554, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. [Text Generation from Knowledge Graphs with Graph Transformers](#). In *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)*, pages 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23*, page 611–626, New York, NY, USA. Association for Computing Machinery.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Shujie Li, Liang Li, Ruiying Geng, Min Yang, Binhua Li, Guanghu Yuan, Wanwei He, Shao Yuan, Can Ma, Fei Huang, and Yongbin Li. 2024. [Unifying structured data as graph for data-to-text pre-training](#). *Transactions of the Association for Computational Linguistics*, 12:210–228.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2024. [A survey on neural data-to-text generation](#). *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. [On LLMs-driven synthetic data generation, curation, and evaluation: A survey](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Sanket Vaibhav Mehta, Jinfeng Rao, Yi Tay, Mihir Kale, Ankur Parikh, and Emma Strubell. 2022. [Improving compositional generalization with self-training for data-to-text generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4205–4219, Dublin, Ireland. Association for Computational Linguistics.
- Pablo Mendes, Max Jakob, and Christian Bizer. 2012. [DBpedia: A multilingual cross-domain knowledge base](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1813–1817, Istanbul, Turkey. European Language Resources Association (ELRA).
- Ali Mousavi, Xin Zhan, He Bai, Peng Shi, Theodoros Rekatsinas, Benjamin Han, Yunyao Li, Jeffrey Pound, Joshua M. Susskind, Natalie Schluter, Ihab F. Ilyas, and Navdeep Jaitly. 2024. [Construction of paired knowledge graph - text datasets informed by cyclic evaluation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3782–3803, Torino, Italia. ELRA and ICCL.
- Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangu Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. [DART: Open-domain structured data record to text generation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.
- Feng Nie, Jinpeng Wang, Jin-Ge Yao, Rong Pan, and Chin-Yew Lin. 2018. [Operation-guided neural networks for high fidelity data-to-text generation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3879–3889, Brussels, Belgium. Association for Computational Linguistics.

- OpenAI et al. 2024. [Gpt-4 technical report](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. [Data-to-text generation with content selection and planning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6908–6915.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. [Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.
- Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scuttheeten, and Patrick Gallinari. 2021. [Data-QuestEval: A referenceless metric for data-to-text semantic evaluation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8029–8036, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. [Investigating pretrained language models for graph-to-text generation](#). In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Martin Vejvar and Yasutaka Fujimoto. 2023. [AS-PIRO: Any-shot structured parsing-error-induced Re-prompting for consistent data-to-text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3550–3563, Singapore. Association for Computational Linguistics.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Luyu Wang, Yujia Li, Ozlem Aslan, and Oriol Vinyals. 2021. [WikiGraphs: A Wikipedia text - knowledge graph paired dataset](#). In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 67–82, Mexico City, Mexico. Association for Computational Linguistics.
- Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. [LLM-powered data augmentation for enhanced cross-lingual performance](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.
- Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. [Challenges in data-to-document generation](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.
- Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. [Learning neural templates for text generation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiannan Xiang, Zhengzhong Liu, Yucheng Zhou, Eric Xing, and Zhiting Hu. 2022. [ASDOT: Any-shot data-to-text generation with pretrained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1886–1899, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *Proceedings of the International Conference on Learning Representations*.

Alex Zhuang, Ge Zhang, Tianyu Zheng, Xinrun Du, Junjie Wang, Weiming Ren, Stephen W. Huang, Jie Fu, Xiang Yue, and Wenhu Chen. 2024. [Structlm: Towards building generalist models for structured knowledge grounding](#).

A Details of Graph Extraction

We utilize the vLLM (Kwon et al., 2023) for efficient inference using GPUs with at least 40 GB of VRAM. We perform graph extraction using nucleus sampling (Holtzman et al., 2020) based decoding strategy with a temperature set to 0.5 and a top-p value of 0.9, which are determined experimentally. Total inference time for processing 6M data samples was approximately 600 hours when utilizing 4 GPUs in parallel. Therefore, we divided the data into divisions that contain 100K samples each to generate the outputs in parallel. Detailed prompts are provided in Table 8.

B Details of Qualitative Analysis

We describe details of human evaluation process and results (§B.1). We then describe GPT-4o evaluation process and prompts (§B.2).

B.1 Details of human evaluation

		GenWiki	TEKGEN	LAGRANGE	WebNLG	WikiOFGraph
1	Triplet	63.52	33.26	24.42	2.25	7.45
	Text	78.38	56.66	37.88	1.26	4.61
2	Triplet	67.19	32.80	16.78	2.38	4.29
	Text	72.83	13.61	5.03	0.96	3.33
3	Triplet	43.00	20.77	10.29	0.00	2.78
	Text	71.80	47.59	28.26	0.23	6.27
4	Triplet	47.28	17.46	30.71	4.16	10.00
	Text	67.61	34.65	28.85	0.80	2.78
5	Triplet	69.63	33.39	12.84	0.00	1.67
	Text	35.00	52.45	35.39	0.25	2.52
Average	Triplet	58.12±12.14	27.54±7.78	19.01±8.44	1.76±1.77	5.24±3.44
	Text	65.12±17.27	40.99±17.4	27.08±13	0.7±0.45	3.9±1.55

Table 6: Individual scores of 5 different human evaluators. The *Average* row includes the mean and standard deviation in the form of *mean±standard deviation*

The human evaluators consist of a diverse group of five individuals from various professional back-

grounds. All evaluators, with their native English proficiency or advanced English language certifications, have prior experience in AI data labeling. We provide the human evaluators with an overview of the guideline and examples to validate graph-text pairs. Evaluators are tasked with assessing whether the target text accurately reflects the information contained within the corresponding triplets. Part of the evaluation guideline⁹ provided with the human evaluators is shown in Figure 8. We give detailed examples and step-by-step instructions to ensure consistency and accuracy to the human evaluators. Some examples of an actual evaluation task completed by the evaluators are shown in Figure 9. Following this process, we collect responses from five human evaluators. The aggregated results, showing the response results for each evaluator, are presented in Table 6.

B.2 Details of GPT-4o evaluation

For the GPT-4o evaluation, we craft a prompt that closely mirrored the guideline provided to the human evaluators, ensuring that the LLM could comprehend and process the task effectively. Using the *gpt-4o-2024-05-13 model*, we evaluated 1,000 samples across five different datasets, resulting in a total cost of \$39.04. The actual prompt used in this evaluation can be found in Table 9. We set the temperature to 1 and allowed it to generate up to 700 tokens. The results to assess the consistency of 5 different datasets are shown in Table 7.

	GenWiki	TEKGEN	LAGRANGE	WebNLG	WikiOFGraph
Triplet	58.74	43.62	34.64	4.93	6.71
Text	48.56	37.37	45.06	6.03	12.23

Table 7: Scores of GPT-4o evaluation.

C Details of Fine-Tuning

In the section 5, we apply bf16 precision, a cosine learning rate decay function with 2k warmup steps, and utilize ZeRO stage 3 optimization through DeepSpeed (Rasley et al., 2020) while employing the AdamW optimizer (Loshchilov and Hutter, 2019). We employ beam search decoding for training and inference, with a temperature set to 1 and beam size set to 5, to ensure a controlled search space exploration while maintaining the diversity of generated outputs. The specific hyperparameters

⁹The full guideline is available at <https://github.com/daehuikim/WikiOFGraph>.

vary depending on the dataset to ensure adequate optimization steps.

For WebNLG (Castro Ferreira et al., 2020), we set the gradient accumulation steps and the number of parallel GPUs to achieve an effective batch size of 8 with a maximum learning rate of 0.3. We evaluate every 4K steps and selected the model with the lowest loss. Total training time was approximately 13 hours using 4 GPUs in parallel (52 GPU hours).

For other datasets, we use an effective batch size of 192 with a maximum learning rate of 0.5. Evaluations are performed every 20k steps, and the model with the minimum loss is selected. Total training time was approximately 70 hours using 4 GPUs in parallel (280 GPU hours).

In the experiments described in section 7, we limit the number of data samples to 50K, similar to the size of WebNLG. Therefore, we conduct the experiments using the same hyperparameter settings as those applied to WebNLG. We select the model based on the one with the lowest loss as determined by evaluations conducted every 4K steps. Total training time was approximately 10 hours using 4 GPUs in parallel (40 GPU hours).

D Details of datasets

In this section, we provide information on the sources of the datasets used in our study and detailed statistics for each dataset.

1. **Wikipedia:** We use Wikipedia from Huggingface datasets. Specifically, we utilize the *20220301.en* split, which can be accessed at [Wikipedia dataset](#).
2. **WebNLG:** We use WebNLG obtained from the official repository. We utilize the English data from version 3.0, which is available at [WebNLG repository](#).
3. **GenWiki:** We obtain GenWiki from [GenWiki repository](#). We use the data found in the ‘fine’ split. Since there is no separate evaluation split, we reserve 10% of the training samples for evaluation purposes.
4. **TekGen:** We obtain TekGen from [Tekgen repository](#). We convert the data into triplets according to the JSON object rules specified in the official repository.
5. **LAGRANGE:** We obtain LAGRANGE from the footnotes of the paper (Mousavi et al., 2024).

We also provide a detailed comparison of the specific statistics and characteristics of each dataset, offering a clear overview of their unique features and differences in Table 10.

[Triplet-Text Pair Validation]

Now, let's explain how to validate Triplet-Text pairs. As mentioned earlier, a triplet includes a subject, predicate, and object. The AI's task involves generating sentences from multiple triplets, making it essential to understand the meaning of each triplet when validating the Triplet-Text pairs.

We will now explain how to determine whether a Triplet-Text pair is correctly or incorrectly aligned, along with some examples.

[Examples]

When you open up a file named as “**number.docx**”, you can see 30 tables per each file. One example is followed.

<Triplet Set>
1. (<S> Glyphostoma epicasta <P> Family <O> Clathurellidae) 2. (<S> Glyphostoma epicasta <P> Class <O> Marine gastropod mollusk) 3. (<S> Glyphostoma epicasta <P> Type <O> Sea snail)
<Reference Text>
Glyphostoma epicasta is a species of sea snail, a marine gastropod mollusk in the family Clathurellidae.
<Error>
[Unused triplets]: x [Unguessable text]: x

Each table contains three distinct fields, described as follows:

<Triplet Set>

This field includes multiple triplets, ranging from 1 to more than 10 triplets. Each triplet is a structured representation of a subject, predicate, and object. Single or multiple triplets together describe a single sentence.

Every triplet has its ID at the beginning(1.(<S> Alice | <P> Occupation | <O> Writer) 1 is ID of this triplet)

<Reference Text>

This field contains the text that corresponds to the <Triplet Set>. The text is generated by understanding the structure and meaning of the triplets in the <Triplet Set>. It should clearly and accurately contain **“Only and Every”** information from <Triplet Set>.

Figure 8: Part of the guidelines for human evaluators.

1.

<Triplet Set>
1.<S> Schruns <P> state <O> Vorarlberg) 2.<S> Schruns <P> district <O> Bludenz) 3.<S> Schruns <P> country <O> Austria)
<Reference Text>
Schruns is in Vorarlberg , the westernmost area of Austria , in the vicinity of Bludenz at 690 meters in the Montafon valley on the Litz river , a tributary of the Ill river .
<Error>
[Unused triplets]: [Unguessable text]:

2.

<Triplet Set>
1.<S> Benedito de Lira <P> birthPlace <O> Alagoas)
<Reference Text>
Benedito de Lira (born May 1 , 1942) is a Brazilian politician . He has represented Alagoas in the Federal Senate since 2011 . Previously , he was a Deputy from Alagoas from 1995 to 1999 .
<Error>
[Unused triplets]: [Unguessable text]:

3.

<Triplet Set>
1.<S> Ian Hornak <P> birthPlace <O> Philadelphia) 2.<S> Ian Hornak <P> nationality <O> American) 3.<S> Ian Hornak <P> birthPlace <O> Pennsylvania)
<Reference Text>
Ian Hornak was born in Philadelphia , Pennsylvania , to parents who immigrated from Slovakia .
<Error>
[Unused triplets]: [Unguessable text]:

Figure 9: Some examples of human evaluation tasks.

Instruction

Your task is to create a set of triplets that can represent all the entities that appear in the given text.

Triplet is consist of three parts (<S>, <P>, <O>).

<S> means subject, <P> means predicate (relation), <O> means object.

You can not simply copy the entities from the text, you need to create a set of triplets that can represent all the entities in the text.

For example, you cannot just use "is" or "are" in <P> part, you need to find a more specific predicate that can represent the relationship between the subject and the object.

Complete the [[TRIPLET]] to represent [[TEXT]], as shown in the Examples.

Please just complete [[TRIPLET]] without saying anything else.

Example 1

[[TEXT]]: The Acharya Institute of Technology is located in Soldevanahalli, Acharya Dr. Sarvapalli Radhakrishnan Road, Hessarghatta Main Road, Bangalore – 560090, Karnataka, India. It is affiliated with the Visvesvaraya Technological University in Belgaum.

[[TRIPLETS]]: (<S>Acharya Institute of Technology|<P>affiliation|<O>Visvesvaraya Technological University), (<S>Visvesvaraya Technological University|<P>city|<O>Belgaum), (<S>Acharya Institute of Technology|<P>state|<O>Karnataka), (<S>Acharya Institute of Technology|<P>country|<O>India), (<S>Acharya Institute of Technology|<P>campus|<O>In Soldevanahalli, Acharya Dr. Sarvapalli Radhakrishnan Road, Hessarghatta Main Road, Bangalore – 560090.)

Example 2

[[TEXT]]: Albert Jennings Fountain was born in Staten Island in New York City and died in Dona Ana County, New Mexico.

[[TRIPLETS]]: (<S>Albert Jennings Fountain|<P>death Place|<O>Doña Ana County, New Mexico), (<S>Albert Jennings Fountain|<P>birth Place|<O>New York City), (<S>Albert Jennings Fountain|<P>birth Place|<O>Staten Island)

Example 3

[[TEXT]]: Abilene, Texas is in Jones County in the United States. Washington, D.C. is the capital of the U.S. with New York City being the largest city. English is their native language.

[[TRIPLETS]]: (<S>United States|<P>capital|<O>Washington, D.C.), (<S>Abilene, Texas|<P>is Part Of|<O>Jones County, Texas), (<S>Jones County, Texas|<P>country|<O>United States), (<S>United States|<P>largest City|<O>New York City), (<S>United States|<P>language|<O>English language)

Query

[[TEXT]]: {text}

[[TRIPLETS]]:

Table 8: Actual prompt used for Graph Extraction. The text in blue indicates where we have inserted the source sentences.

<p>Instruction</p> <p>Your job is to validate Triplet-Text pair data for Triplet-to-Text generation task. Based on the provided instructions and examples, write all errors in the Triplet-Text pairs present in the given query.</p> <p>Triplet Composition:</p> <p>Format: (<S>subject <P>predicate <O>object)</p> <p>A triplet is composed of three elements: subject, predicate, and object. Each of these elements is delineated by specific symbols with capital letters such as <S>, <P>, and <O>. These symbols help to clearly identify the role of each element within the triplet.</p> <ul style="list-style-type: none"> • <S> stands for the subject, which represents the entity that is performing an action or being described. • <P> stands for the predicate, which illustrates the relationship or action that connects the subject to the object. • <O> stands for the object, which is the entity that is receiving the action or being described in relation to the subject. <p>Error Types:</p> <p>A. Unused Triplet {Example} {Explanation}</p> <p>B. Unguessable Text {Example} {Explanation}</p> <p>Please just fill the format below. No need to write any extra explanations.</p> <p>Query</p> <p>[[TRIPLET SET]]: {triplet}</p> <p>[[TEXT]]: {text}</p> <p>[[ERRORS]]:</p> <ul style="list-style-type: none"> • [Unused Triplets]: • [Unguessable Text]:
--

Table 9: Actual prompt for validating triplet-text pairs. {Example} and {Explanation} are omitted due to the length of the prompt. The text in blue indicates where we have inserted the examples.

	WebNLG	TekGen	GenWiki	LAGRANGE	WikiOFGraph (Ours)
# of samples	35,426	6,310,061	681,436	3,075,058	5,851,776
# of unique predicate	372	50,861	287	1,167	140,733
# of unique entity	3,211	4,249,337	866,373	2,904,407	8,217,819
# of triples (m/M/avg)	1/7/2.96	1/54/1.73	1/10/2.64	1/135/4.02	1/17/3.62
Human Annotator	o	x	x	x	x
Domain Diversity	↓	↑	↑	↑	↑
Characteristic	Crowdsourcing	Distant supervision	For unsupervised learning	Second-hop matching, Triple Augmentation	LLM graph extraction
Ontology	DBpedia	Wikidata	DBpedia	Wikidata	x

Table 10: Detailed comparison of dataset statistics and characteristics.