

# Evaluating Generative Models for Graph-to-Text Generation

Shuzhou Yuan, Michael Färber

Karlsruhe Institute of Technology (KIT)

{shuzhou.yuan, michael.farber}@kit.edu

## Abstract

Large language models (LLMs) have been widely employed for graph-to-text generation tasks. However, the process of finetuning LLMs requires significant training resources and annotation work. In this paper, we explore the capability of generative models to generate descriptive text from graph data in a zero-shot setting. Specifically, we evaluate GPT-3 and ChatGPT on two graph-to-text datasets and compare their performance with that of finetuned LLM models such as T5 and BART. Our results demonstrate that generative models are capable of generating fluent and coherent text, achieving BLEU scores of 10.57 and 11.08 for the AGENDA and WebNLG datasets, respectively. However, our error analysis reveals that generative models still struggle with understanding the semantic relations between entities, and they also tend to generate text with hallucinations or irrelevant information. As a part of error analysis, we utilize BERT to detect machine-generated text and achieve high macro-F1 scores. We have made the text generated by generative models publicly available.<sup>1</sup>

## 1 Introduction

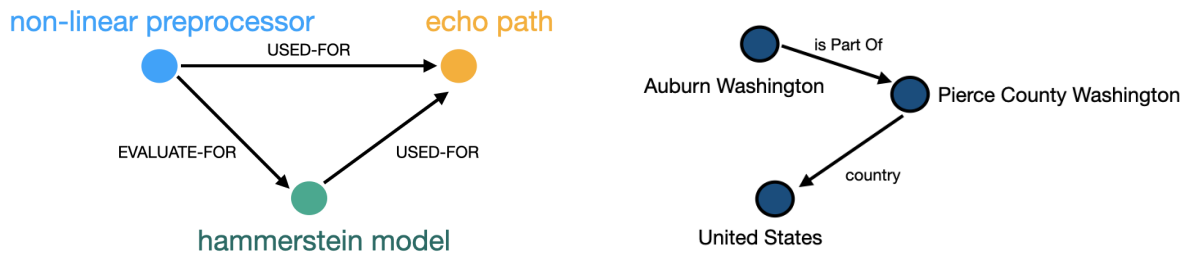
Graph-to-text generation is a subtask of data-to-text generation and Natural Language Generation (NLG) (Gatt and Krahmer, 2018). Its purpose is to generate fluent descriptive text based on the structure of a given graph (see Figure 1). With the widespread use of graph structured data, this technique plays a crucial role in various Natural Language Processing applications, including question answering, dialogue systems, and data augmentation (He et al., 2017; Zhao et al., 2020; Josifoski et al., 2023). Previous research on model architectures has achieved significant performance

on graph-to-text generation benchmarks (Koncel-Kedziorski et al., 2019; Ribeiro et al., 2020; Zhao et al., 2020; Li et al., 2021; Ribeiro et al., 2021b). In particular, Ribeiro et al. (2021a) achieved state-of-the-art performance by employing large pre-trained language models and sufficient training data. However, the zero-shot setting for graph-to-text generation remains challenging due to the inconsistent input format (unstructured text vs. preformatted text) between pretraining and fine-tuning stages for large language models.

Recently, generative models such as GPT-3 (Brown et al., 2020), InstructGPT (Ouyang et al., 2022), and ChatGPT have gained tremendous attention in both the NLP research community and the general public. Researchers have evaluated these models on various NLP benchmarks in the zero-shot setting (Bang et al., 2023; Jiao et al., 2023; Ahuja et al., 2023). However, their ability to process structured data, and in particular graph data, such as knowledge graphs, is understudied and worth being explored (Bang et al., 2023). Given the significant resources and annotations required for training graph-to-text generation models (Li et al., 2021), utilizing a zero-shot setting could save training resources and prove advantageous for both economic and ecological reasons.

Previous approaches has come up with a neural pipeline to enable zero-shot for graph-to-text generation but didn't use generative models (Kasner and Dusek, 2022). In contrast, our approach adopts the zero-shot setting by using prompts as instructions for generative models, specifically GPT-3 and ChatGPT (Brown et al., 2020; Ouyang et al., 2022). We evaluate the models' ability to translate graph data into fluent text using the test sets from two widely used graph-to-text generation datasets: WebNLG (Gardent et al., 2017) and AGENDA (Koncel-Kedziorski et al., 2019). Following the method of Ribeiro et al. (2021a), we represent the

<sup>1</sup>[https://github.com/ShuzhouYuan/Eval\\_G2T\\_GenModels](https://github.com/ShuzhouYuan/Eval_G2T_GenModels)



(a) Generate paper abstract from title, entities and graph: **<title>** Significance-aware Hammerstein group models for non-linear acoustic echo cancellation. **<entities>** non-linear preprocessor echo path hammerstein model **<graph>** **<H>** non-linear preprocessor **<R>** USED-FOR **<T>** echo path **<H>** preprocessor **<R>** EVALUATE-FOR **<T>** hammerstein model **<H>** hammerstein model **<R>** USED-FOR **<T>** echo path

(b) Generate text from graph: **<H>** Auburn Washington **<R>** is Part Of **<T>** Pierce County Washington **<H>** Pierce County Washington **<R>** country **<T>** United States

Figure 1: Examples of graph structures, prompts and linearized graphs of (a) AGENDA and (b) WebNLG.

graph as a linearized sequence of text for input to the models (see Figure 1).

To assess the performance of the generative models, we conduct a comprehensive evaluation on each dataset. Employing machine translation metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) to the generated texts, we reveal that the generative models fall short of matching the quality achieved by state-of-the-art approaches. To identify patterns of mistakes made by the generative models, we perform error analysis by comparing the generated texts with the reference texts. Additionally, we fine-tune a BERT model to detect the machine-generated text. We make the texts generated by the models available on GitHub to facilitate future research on the analysis of machine-generated text and trustworthy AI.

In summary, our study aims to assess the performance of generative models in the zero-shot setting for graph-to-text generation using two distinct benchmarks. Our contribution lies in conducting a rigorous quantitative analysis of the results, shedding light on the effectiveness of generative models in this domain.

## 2 Related Work

**Graph-to-text generation.** Various efforts have been made to enhance graph-to-text generation using neural network models. They can be categorized into two main types: Graph Neural Network (GNN) based models and Language Model (LM) based models. GNN-based models typically employ a graph encoder to encode the graph struc-

ture (Beck et al., 2018; Marcheggiani and Perez-Beltrachini, 2018; Damonte and Cohen, 2019; Koncel-Kedziorski et al., 2019; Ribeiro et al., 2019; Li et al., 2021). In contrast, LM-based models do not rely on the graph structure but purely on the sequence of tokens in the text. As such, graphs have first been transformed into a linearized representation before being fed into LMs to generate coherent text (Harkous et al., 2020; Ribeiro et al., 2021a,b). Besides GNN and LM, previous works have also explored the use of Recurrent Neural Networks (RNNs) such as LSTM and GRU for graph-to-text generation (Song et al., 2018; Zhao et al., 2020; Guo et al., 2020). We follow the approach of Konstantas et al. (2017) and other prior works by using a linearized graph as input for generative models.

**Generative Models.** Generative language models, such as GPT-3 (Brown et al., 2020), InstructGPT (Ouyang et al., 2022), and ChatGPT, have been designed to learn and generate natural language text. These models are based on the transformer decoder architecture (Vaswani et al., 2017), which enables them to handle large amounts of training data and perform zero-shot applications. While GPT-3 has made a significant breakthrough in text completion, InstructGPT and ChatGPT possess unique characteristics that align user intent with a conversational style. These models are trained using supervised fine-tuning and reward modeling, allowing them to generate high-quality responses that accurately reflect the user’s needs and preferences. InstructGPT and ChatGPT are first fine-tuned on the GPT-3 model through supervised learning and then further trained using reinforcement learning based on human feedback.

	AGENDA	WebNLG
Number of Instance	1,000	1,862
Average Input Tokens	169	66

Table 1: Statistics of test sets from AGENDA and WebNLG.

As demonstrated by Ouyang et al. (2022), this approach substantially improves the model’s performance on NLP benchmarks. Although there have been numerous reports and research evaluating the performance of generative models in various NLP applications such as summarization (Bang et al., 2023), machine translation (Jiao et al., 2023), and multilingual evaluation (Ahuja et al., 2023), our work focuses on the generative models’ capability to handle structured data.

### 3 Dataset

We evaluate generative models using the AGENDA and WebNLG datasets, as they are widely used in recent research on graph-to-text generation (Koncel-Kedziorski et al., 2019; Ribeiro et al., 2021a; Li et al., 2021) and as they represent different domains: scholarly domain and general domain (e.g., as given in Wikipedia). We focus on the test sets of AGENDA and WebNLG for our experiments, as the models do not require further training. In the following, we briefly describe the used datasets.

**AGENDA.** Abstract GENERation DATaset (AGENDA) is a dataset that pairs knowledge graphs with paper abstracts from scientific domains (Koncel-Kedziorski et al., 2019). The graphs in AGENDA were automatically extracted from the SciIE information extraction system (Luan et al., 2018). Each instance in AGENDA includes the title, entities, graph, and abstract of a paper. We use the title, entities, and graph as input for the models.

**WebNLG.** This dataset is a benchmark for mapping sets of RDF triples to text (Gardent et al., 2017). The RDF triples are subgraphs of the knowledge graph DBpedia (Auer et al., 2007), while the texts describe the graphs in one or a few sentences. The WebNLG challenge<sup>2</sup> has released several versions of this dataset since 2017. In order to compare with previous work, we take the test data of

<sup>2</sup><https://synalp.gitlabpages.inria.fr/webnlg-challenge/>

WebNLG challenge 2017 for our experiments.

## 4 Experiments

**Data Preprocessing.** Since GPT-3 and ChatGPT require a sequence of text as input, we convert the graph structure into a linearized representation following Ribeiro et al. (2021a). To assist the models in identifying the head, relation, and tail entities, we prepend <H>, <R>, and <T> tokens before the entities, as done in previous work (Harkous et al., 2020). In the AGENDA dataset, each sample also includes a title and entities. Thus, we additionally add <title>, <entities>, and <graph> tokens (see Figure 1).

**Model Settings.** We use the GPT-3 model variant `text-davinci-003` and the ChatGPT model variant `gpt-3.5-turbo-0301` for our experiments. Each instance is treated as a single request, and the first response from the model is taken as the generated text. The prompt used for the models plays a significant role as it serves as the task description and directly influences the content of the generated text. Previous work designed prompts by asking ChatGPT (Jiao et al., 2023). Following their approach, we ask ChatGPT to provide prompts: “Please provide prompts or templates for graph-to-text generation:”. Since AGENDA and WebNLG have different data structures, we use the prompt “Generate paper abstract from title, entities, and graph:” for AGENDA. For WebNLG, we use the prompt “Generate text from graph:”. We expect that in this way the generated text fits the format of a scientific paper abstract better for AGENDA, while the models generate texts in open domain for WebNLG.

**Baseline.** Similar to our experimental methodology, Ribeiro et al. (2021a) finetuned T5 and BART using linearized graphs as input and generated descriptive texts. Therefore, we consider their findings as the baseline for comparison with our own experiments.

**Evaluation.** Following related work, we implement a thorough evaluation with metrics BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), RougeL (Lin, 2004) and Chrf++ (Popović, 2017). Additionally, to assess the semantic meaning and coherence of the generated text, we employ BLEURT (Sellam et al., 2020), a metric that evaluates not only the surface match of n-grams but also the semantic representation extracted from a pretrained BERT (Devlin et al., 2019) model.

Model	BLEU $\uparrow$	METEOR $\uparrow$	RougeL $\uparrow$	Chrf++ $\uparrow$	BLEURT $\uparrow$
T5 <sub>large</sub> (Ribeiro et al., 2021a)	22.15	23.73	-	-	-13.96
BART <sub>large</sub> (Ribeiro et al., 2021a)	<b>23.65</b>	<b>25.19</b>	-	-	<b>-10.93</b>
GPT-3	8.34	14.88	24.99	41.42	-32.54
ChatGPT	10.57	17.02	<b>25.22</b>	<b>45.86</b>	-28.05

Table 2: Results on AGENDA.

Dataset	BLEU $\uparrow$	METEOR $\uparrow$	RougeL $\uparrow$	Chrf++ $\uparrow$	BLEURT $\uparrow$
T5 <sub>large</sub> (Ribeiro et al., 2021a)	<b>59.70</b>	<b>44.18</b>	-	<b>75.40</b>	-
BART <sub>large</sub> (Ribeiro et al., 2021a)	54.72	42.23	-	72.29	-
GPT-3	20.36	26.95	<b>45.64</b>	57.95	<b>13.39</b>
ChatGPT	11.08	23.89	35.87	48.75	-10.99

Table 3: Results on WebNLG.

## 4.1 Results

Our results are summarized in Table 2 and 3. As comparison, we take the results from Ribeiro et al. (2021a), which are achieved by finetuned BART and T5.

The results obtained from AGENDA demonstrate that finetuned BART and T5 models outperform generative models in terms of state-of-the-art performance. Both T5 and BART achieve BLEU scores exceeding 20, while GPT-3 only attains a BLEU score of 8.34 and ChatGPT achieves 10.57. Consistently, other evaluation metrics align with the BLEU scores, further highlighting the limited performance of generative models without fine-tuning. Notably, ChatGPT exhibits a slightly improved performance compared to GPT-3 on the AGENDA benchmark. Analysis of the results reveals that ChatGPT consistently outperforms GPT-3 across all metrics, showcasing a 2.23 higher BLEU score, a 2.14 higher METEOR score, a 0.23 higher RougeL score, a 4.44 higher Chrf++ score, and a 4.49 higher BLEURT score.

Examining the results from WebNLG, it becomes evident that fine-tuned T5 and BART models consistently outperform generative models without fine-tuning. Notably, both T5 and BART achieve BLEU scores exceeding 50, whereas generative models only attain a BLEU score of 11.08 for ChatGPT and 20.36 for GPT-3. Surprisingly, GPT-3 outperforms ChatGPT on the WebNLG benchmark with a BLEU score that is 9.28 higher, a METEOR score that is 3.06 higher, a RougeL score that is 9.77 higher, and a Chrf++ score that is 9.20 higher. The primary reason for this difference is that ChatGPT

tends to produce hallucinations easily and generates longer text. We provide further elaboration on two examples in Section 5.

## 5 Error Analysis

We observe that the texts generated by generative models contain errors following similar patterns. In Table 4 and Table 5, we show two examples from AGENDA and WebNLG.

As shown in the example of Table 4, generative models are capable of generating fluent paper abstracts based on the provided input, but they contain errors that contradict the facts represented by the graph structure. For instance, the triple “accuracy <R> EVALUATE-FOR <T> correlation analysis” in the example graph is wrongly generated by GPT-3 as “which (correlation analysis) is then used to evaluate the estimation accuracies”. Also, ChatGPT describes it as “(The authors) apply clustering and correlation analysis to evaluate the accuracy”. In both cases, the models reverse the relationship between “accuracy” and “correlation analysis”.

Furthermore, ChatGPT ignores the relations “PART-OF” and “USED-FOR” in the input. This suggests that generative models still face challenges in comprehending the semantic relations between entities, which can result in the generation of inaccurate text.

While generating paper abstracts is complex and challenging, generating short descriptive text from a knowledge graph is relatively more straightforward. As LLMs, GPT-3 and ChatGPT are trained on large corpora and thus already contain world knowledge. In the example provided in Table 5,



AGENDA	Example
Input	Generate paper abstract from title, entities and graph: <title> Estimation of semantic case of Japanese dialogue by use of distance derived from statistics of dependency . <entities> statistics of dependent noun-particle-verb triples atr dialogue corpus measure of distance semantic cases noun-particle-verb triples consistency rates single-case clusters clustering analysis correlation analysis estimation accuracies accuracy <graph> <H> statistics of dependent noun-particle-verb triples <R> USED-FOR <T> measure of distance <H> noun-particle-verb triples <R> PART-OF <T> atr dialogue corpus <H> semantic cases <R> PART-OF <T> atr dialogue corpus <H> accuracy <R> EVALUATE-FOR <T> correlation analysis
Reference	In an attempt to estimate the semantic cases for noun-particle-verb triples in the atr dialogue corpus, the authors propose a measure of distance based on statistics of dependent noun-particle-verb triples. A clustering analysis of all the triples in the corpus was conducted using the measure of distance. Competence of the proposed measure of distance is verified by examination of the distribution of the single-case clusters. By use of the score derived from the measure of distance of the training corpus, the authors conducted the estimation of the correct semantic case for a given noun-particle-verb triples in the test corpus. The result remarkably differentiates the particles with respect to the estimation accuracies. For instance, particle ‘wo’ has accuracies over 80 %, while ‘de’ has accuracies less than 40%. The correlation analysis between the accuracy and the consistency rates indicates that the particles of higher consistency have also tendencies to higher accuracies.
GPT-3	This paper aims to estimate semantic cases of Japanese dialogue using a measure of distance derived from statistics of dependent noun-particle-verb triples from the ATR dialogue corpus. Clustering analysis and <b>correlation analysis</b> are employed to examine the consistency rates of single-case clusters, which is then used <b>to evaluate</b> the estimation <b>accuracies</b> of the distance measure.
ChatGPT	This paper proposes a method to estimate the semantic case of Japanese dialogue using a measure of distance derived from the statistics of dependent noun-particle-verb triples. The authors use the ATR dialogue corpus and apply clustering and <b>correlation analysis to evaluate the accuracy</b> of their method. They report consistent rates for single-case clusters and show that their method achieves high estimation accuracies. The study demonstrates the potential of using statistics of dependency for analyzing Japanese dialogue.

Table 4: Paper abstract generated by GPT-3 and ChatGPT, incorrect texts are marked with red.

generative models generate descriptive text fluently based on the structured input. However, GPT-3 produces text with incorrect facts. For instance, “Abner W. Sibal” is the name of a person, but it is recognized as the name of a battle by GPT-3. While GPT-3 produces text with incorrect facts, it is worth noting that ChatGPT-generated text not only covers the input information, but also contains redundant messages from its internal knowledge (hallucinations). Furthermore, the sentence “Ab-

ner W. Sibal was a prominent figure who faced great challenges” generated by ChatGPT has made a subjective judgement about the character and may cause unnecessary bias to potential users.

To investigate the difference between model-generated texts and human-written reference texts, we create datasets containing both types of text and finetune a pretrained BERT model for a binary text classification task. The statistics of our datasets are presented in Table 6.

WebNLG	Example
Input	Generate text from graph: <H> Abner W. Sibal <R> battles <T> World War II <H> World War II <R> commander <T> Joseph Stalin
Reference	Joseph Stalin was a commander in World War II in which Abner W Sibal also fought.
GPT-3	Joseph Stalin was the commander of World War II and <b>Abner W. Sibal was one of the battles</b> fought during the war.
ChatGPT	Abner W. Sibal was a prominent figure who faced great challenges during his time as a soldier in World War II. <b>He fought bravely as a commander, leading his troops with dedication and determination. His efforts were critical to the success of the war effort and his name became synonymous with bravery and patriotism. During World War II, the world was gripped in the fight against tyranny and oppression. Abner W. Sibal was one of the many who answered the call to serve his country. Through his leadership skills and strategic thinking, he was able to navigate the treacherous waters of war and emerge victorious. However, the fight against the Axis powers was not just a battle against military might. It was a battle against the ideologies that fueled the enemy’s aggression. As a commander, Sibal had to navigate these complexities and ensure that his troops remained committed to the cause of freedom. Throughout the war, Sibal faced many challenges and obstacles. But he remained steadfast and only grew more determined to see the war through to its ultimate conclusion. His efforts were recognized and honored, not just by his own country but also by leaders around the world. Even the likes of Joseph Stalin, renowned commander and leader of the Soviet Union, recognized the critical role Sibal played in the victory against fascism. Today, Abner W. Sibal’s name remains a testament to the bravery and dedication of those who fought in World War II. His legacy serves as an inspiration to future generations, reminding them of the importance of standing up against tyranny and oppression wherever it may arise.</b>

Table 5: Descriptive text generated by GPT-3 and ChatGPT, incorrect texts are marked with **red**, hallucinations are marked with **blue**.

Dataset	Machine-generated instance	Human-written instance
AGENDA	1000	1000
WebNLG	1862	4894
All	2862	5894

Table 6: Statistics of ChatGPT generated datasets

We create several datasets for AGENDA, WebNLG, and a combined dataset containing both AGENDA and WebNLG examples. The training and test sets are split in an 80:20 ratio. We fine-tune BERT for five epochs using the AdamW optimizer (Loshchilov and Hutter, 2019). As shown in Table 7, BERT achieves high scores across all datasets. This demonstrates that generative models generate text that follows similar patterns, and a state-of-the-art text classifier can easily distinguish between

them.

## 6 Conclusion

In this paper, we explored the capabilities of generative models in generating coherent text from structured data, focusing on two benchmarks: AGENDA and WebNLG. To achieve this, we adopted the linearized graph representation approach employed in prior work. Leveraging the zero-shot ability of language models, we prepended the prompt to the

Model	Accuracy	Macro F1
GPT-3 <sub>AGENDA</sub>	98.00	98.00
ChatGPT <sub>AGENDA</sub>	100	100
GPT-3 <sub>WebNLG</sub>	91.64	89.25
ChatGPT <sub>WebNLG</sub>	96.82	95.75
GPT-3 <sub>All</sub>	93.55	92.38
ChatGPT <sub>All</sub>	96.40	95.82

Table 7: Results of BERT to detect GPT-3 and ChatGPT generated text.

input text as an instruction for both GPT-3 and ChatGPT. We conducted a comprehensive evaluation using various metrics. Our findings reveal that generative models fall short of surpassing previous models that have been trained and finetuned on large volumes of training data. These results highlight the limitations of generative models in achieving state-of-the-art performance in graph-to-text generation tasks.

Furthermore, we conducted an error analysis of the text generated by the models. The generative models struggle in capturing the relationships between entities and often produce unrelated information, leading to hallucinations. To further investigate the machine generated text, we employ finetuned BERT to conduct a text classification task. BERT achieves high F1 scores in distinguishing between machine-generated text and human-written text. Our study provides extensive evaluation of generative models for graph-to-text generation. Future work should focus on refining machine-generated text and reducing hallucinations for graph-to-text generation by using generative models.

## 7 Ethical Consideration and Limitation

We observe that generative models may generate text containing fake facts or offensive content. And the datasets we collected may also contain incorrect or offensive statements. We do not support the views expressed in the machine generated text, we merely venture to analyze the machine generated text and provide an useful resource for future research.

As the limitation of this work, we found out that the reproducibility of GPT-3 and ChatGPT is questionable. The models often return different response from same request, which makes our results

hard to reproduce and the brings randomness to the evaluation scores.

## Acknowledgements

We thank the anonymous reviewers for their helpful comments. We also would like to thank Nicholas Popovic for his feedback on this work.

## References

- Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.
- 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.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multi-task, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- 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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Marco Damonte and Shay B. Cohen. 2019. **Structural neural encoders for AMR-to-text generation**. 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 3649–3658, Minneapolis, Minnesota. Association for Computational Linguistics.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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. *Journal of Artificial Intelligence Research*, 61:65–170.
- Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020. [CycleGT: Unsupervised graph-to-text and text-to-graph generation via cycle training](#). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 77–88, Dublin, Ireland (Virtual). Association for Computational Linguistics.
- Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. [Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. 2017. [Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 199–208, Vancouver, Canada. Association for Computational Linguistics.
- Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.
- Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. *arXiv preprint arXiv:2303.04132*.
- 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.
- 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.
- Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. [Neural AMR: Sequence-to-sequence models for parsing and generation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.
- Junyi Li, Tianyi Tang, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2021. [Few-shot knowledge graph-to-text generation with pre-trained language models](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1558–1568, Online. Association for Computational Linguistics.
- 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.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations*, ICLR’19.
- Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. [Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.
- Diego Marcheggiani and Laura Perez-Beltrachini. 2018. [Deep graph convolutional encoders for structured data to text generation](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 1–9, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- 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. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Leonardo F. R. Ribeiro, Claire Gardent, and Iryna Gurevych. 2019. [Enhancing AMR-to-text generation with dual graph representations](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3183–3194, Hong Kong, China. Association for Computational Linguistics.
- Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021a. [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.
- Leonardo F. R. Ribeiro, Yue Zhang, Claire Gardent, and Iryna Gurevych. 2020. [Modeling global and local node contexts for text generation from knowledge graphs](#). *Transactions of the Association for Computational Linguistics*, 8:589–604.
- Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021b. [Structural adapters in pretrained language models for AMR-to-Text generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4269–4282, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. [Bleurt: Learning robust metrics for text generation](#). In *ACL*.
- Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. [A graph-to-sequence model for AMR-to-text generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.
- 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.
- Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. [Bridging the structural gap between encoding and decoding for data-to-text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.