Generating Faithful Text From a Knowledge Graph with Noisy Reference Text

Tahsina Hashem[1], **Weiqing Wang**[1], **Derry Tanti Wijaya**[2],
Mohammed Eunus Ali[3], **Yuan-Fang Li**[1]
[1]Department of Data Science & AI, Monash University, Australia
[2]Department of Data Science, Monash University, Indonesia
[3]Department of CSE, Bangladesh University of Engineering and Technology, Bangladesh
{tahsina.hashem, Teresa.Wang, derry.wijaya, yuanfang.li}@monash.edu;
eunus@cse.buet.ac.bd

Abstract

Knowledge Graph (KG)-to-Text generation aims at generating fluent natural-language text that accurately represents the information of a given knowledge graph. While significant progress has been made in this task by exploiting the power of pre-trained language models (PLMs) with appropriate graph structure-aware modules, existing models still fall short of generating faithful text, especially when the ground-truth natural-language text contains additional information that is not present in the graph. In this paper, we develop a KG-to-text generation model that can generate faithful natural-language text from a given graph, in the presence of noisy reference text. Our framework incorporates two core ideas: Firstly, we utilize contrastive learning to enhance the model's ability to differentiate between faithful and hallucinated information in the text, thereby encouraging the decoder to generate text that aligns with the input graph. Secondly, we empower the decoder to control the level of hallucination in the generated text by employing a controllable text generation technique. We evaluate our model's performance through the standard quantitative metrics as well as a ChatGPT-based quantitative and qualitative analysis. Our evaluation demonstrates the superior performance of our model over state-of-the-art KG-to-text models on faithfulness.

1 Introduction

A knowledge graph (KG) is a structured representation of information as a network of interconnected real-world entities, and relationships. The task of KG-to-text generation has been proposed (Ribeiro et al., 2020a; Koncel-Kedziorski et al., 2019) to make this structured information more accessible to humans, aiming to generate fluent, informative, and faithful natural-language sentences that should describe the contents of an input KG. Recently, this task plays a significant role in a variety of applications such as knowledge-grounded dialogue generation (Zhou et al., 2018; Zhao et al., 2020), story generation (Guan et al., 2019; Ji et al., 2020), event narration (Colas et al., 2021a), and question-answering (Agarwal et al., 2021; Chen et al., 2023; Saxena et al., 2020).

Significant progress has been made in the KG-to-text generation task by utilizing a set of Transformer-based (Vaswani et al., 2017) pre-trained language models (PLMs) such as BART (Lewis et al., 2019), T5 (Raffel et al., 2020) or GPT (Radford et al., 2019) with appropriate graph structure-aware modules (Ke et al., 2021; Colas et al., 2022; Han and Shareghi, 2022). However, ensuring the faithfulness of KG-to-text generation, i.e. reducing hallucinations (Ji et al., 2022; Wang et al., 2022; Raunak et al., 2021; Rebuffel et al., 2022), is an under-explored problem, and existing KG-to-text models fall short of generating faithful text when the ground-truth text of the training dataset contains wrong or extra information that is not consistent with the input.

Figure 1 shows an example of a small KG about a house, which contains information on its internal features and neighborhood, and the corresponding ground-truth reference text, from a real-world real-estate KG (Das et al., 2021). The ground-truth text, while summarizing the features of the house accurately, also mentions some information that is not available in the input KG (i.e. extrinsic hallucination, highlighted in red).

When a KG-to-text model is trained with such hallucinated reference text, it is likely to produce text that is also hallucinated. This hallucination problem significantly reduces the faithfulness and thus trustworthiness of the generated text. Thus, the ability to reduce hallucination in the presence of noisy reference text is important for the practi-

Proceedings of the 16th International Natural Language Generation Conference, pages 106–122
September 11–15, 2023.

cal application of KG-to-text and other NLG techniques, especially in mission- and safety-critical domains such as medical diagnostics and scientific research.

A number of techniques have been proposed (Ji et al., 2022) to control this hallucination problem in abstractive summarization, table-to-text generation, generative question-answering, neural machine translation, and knowledge-grounded dialogue generation (Wang et al., 2022; Tang et al., 2022; Rebuffel et al., 2022; Krishna et al., 2021; Zhou et al., 2021; Zhang et al., 2022). However, to the best of our knowledge, controlling hallucination in graph-to-text generation with noisy reference text has not been investigated.

In this paper, we propose a novel framework to address this important and practical problem. Our framework combines contrastive learning technique and controllable text generation. Contrastive learning enables the model to distinguish between faithful and hallucinated text and guides the decoder to generate faithful text instead of hallucinated text. The controllable text generation technique learns the level of hallucination from noisy training text and controls (i.e. minimizes) the level of hallucinated information in the generated text. Our framework can be employed in any KG-to-Text encoder-decoder model to generate faithful natural language text from a given KG, in the presence of noisy reference text.

Our contributions are as follows:

- We propose a framework to deal with the hallucination problem in KG-to-text generation task. Our framework comprises two core ideas: (i) Employing contrastive learning to enable the KG-to-text generation model to better differentiate between faithful and hallucinated information in the reference text and guide the decoder to generate text that is faithful to KG. (ii) Controlling the level of hallucination while generating text from KG using a controllable text generation technique.
- We conduct experiments and evaluate performance using a standard quantitative analysis with automatic metrics. Our comprehensive evaluation on two noisy datasets demonstrates the superior performance of our proposed model over the state-of-art KG-to-text generation models on faithfulness metrics.
- We further propose and perform novel ChatGPT-based quantitative and qualitative evaluations to assess the performance of our model more comprehensively. The evaluation also shows our model's effectiveness in generating faithful text over existing KG-to-text generation models.

2 Related Work

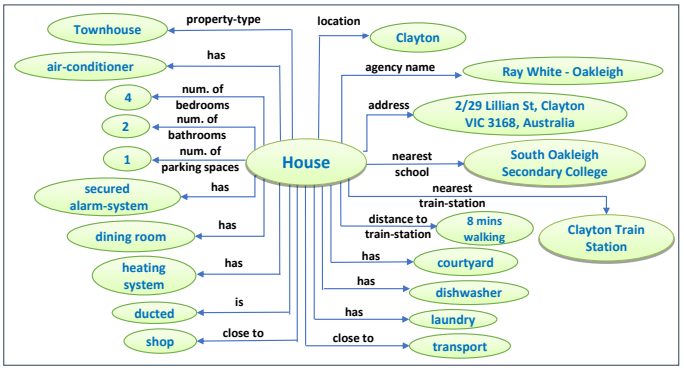
2.1 Knowledge Graph-to-Text Generation

KG-to-text generation techniques (Koncel-Kedziorski et al., 2019; Guo et al., 2020; Ribeiro et al., 2020b; Chen et al., 2020) utilize graph neural networks (Veličković et al.) and graph Transformers (Vaswani et al., 2017) to effectively encode a graph's structural information. With the rapid advancement of pre-trained language models (PLMs) (Lewis et al., 2019; Raffel et al., 2020; Radford et al., 2019), researchers have started adapting and fine-tuning these models to KG-to-text generation tasks and obtained better results compared to previous models (Ribeiro et al., 2021; Chen et al., 2020; Kale and Rastogi, 2020). Recently, researchers further improved the KG-to-text models' performance by integrating pre-trained language models with appropriate graph-structure-aware modules (Ke et al., 2021; Colas et al., 2022) and employing some graph masking pre-training tasks (Ke et al., 2021; Han and Shareghi, 2022).

However, we have empirically observed that although these state-of-art KG-to-text generation models (Ke et al., 2021; Colas et al., 2022; Han and Shareghi, 2022) introduce graph aware encoders and/or apply graph masking pre-training strategies to enhance graph-text alignments, still these models are struggling with hallucination problems when trained with noisy input ground-truth text.

2.2 Controlling Hallucinations in Text Generation

This hallucination problem is well explored in other natural language generation tasks such as in table-to-text generation, summarization, dialogue generation, question-answering, and neural machine translation. Planning (Su et al., 2021) or skeleton-based method (Wang et al., 2021), joint learning strategy (Xu et al., 2021), Bayes training framework (Tian et al., 2019), table-text optimal-transport matching strategy (Wang et al., 2020), control token approach (Filippova, 2020) are widely used in controlling hallucinations in table-to-text generation tasks. Most recently, Rebuffel et



(a) House knowledge graph

Exceptional Modern Elegance! Luxury living in top location!: Fantastic opportunity available for the first-home buyer or astute investor to secure this beautiful double storey residence situated within few minute walk to Monash Medical Centre and Clayton train station, and minutes away from Clayton Shopping Centre, Clayton Community Centre and Monash University, you will not find a better growth location for this exclusive masterpiece. This architecturally-designed 4 bedrooms townhouse, with spacious living room and dining area, large open-plan U-shaped kitchen, downstairs huge master bedroom with walk-in-robe and superb ensuite come with spa bath, and upstairs three other large bedrooms all with built-in-robes, separate retreat areas, central bathroom and a separate toilet. Comprises: 900mm cook top and range hood, two drawer dishwasher, Jarrah Hardwood staircase and flooring, large separate laundry room, separate powder room, ducted heating, separate air-conditioning, single lock-up garage with internal access, security alarm system, private rear courtyard.

(b) Ground-Truth Text

Figure 1: A sample knowledge graph for the House dataset with its ground-truth text. The red colored text in the ground-truth text represents extrinsic hallucination information.

al. (Rebuffel et al., 2022) proposed a multi-branch decoder approach to control hallucination at decoding time in this area.

Prior works have also focused on minimizing hallucinations in summarization, dialogue generation, question-answering and neural machine translation areas. Some of the recent hallucination mitigation techniques are based on control token approach (Filippova, 2020; Rashkin et al., 2021; Wang et al., 2022), contrastive learning approach (Cao and Wang, 2021; Tang et al., 2022), generate then-refine strategy (Dziri et al., 2021), a routing transformer based approach (Krishna et al., 2021) and self-training of neural machine translation based approach (Zhou et al., 2021). To the best of our knowledge, no work has been done in graph-to-text generation tasks with hallucinated ground-truth text.

2.3 Evaluation using ChatGPT

Large language models such as ChatGPT have recently been employed for evaluating the quality and factual consistency of the generated text in NLP tasks with respect to the source input through ranking, rating, and entailment inference (Kocmi and Federmann, 2023; Wang et al., 2023; Luo et al., 2023). Luo et al. (2023) closely investigated ChatGPT's ability under a zero-shot setting with three factual consistency evaluation tasks: binary entailment inference, summary ranking, and consistency rating. Experimental findings show that ChatGPT generally performs better than previous evaluation metrics across the three tasks, demonstrating its significant potential for factual consistency evaluation. However, they also point out some limitations of ChatGPT such as its preference on lexical similarity instead of semantic entailment, false reasoning, and poor understanding of instructions. Moreover, while these approaches can compute an overall faithfulness score of the output text, they fall short in terms of explaining the score e.g., by quantifying the amount of hallucination (out of all the output facts, how many are hallucinated?), precision (out of all the output facts, how many are input facts?) and recall (out of all the input facts, how many appear in the output?). In this work, we use ChatGPT to quantify each of these values and obtain a finer-grained explanation of what a faithfulness score entails.

3 Proposed Model

3.1 Problem Formulation

Let $G = (V, E)$ represent a knowledge graph, where $V = \{e_1, e_2, \ldots, e_{|V|}\}$ represents the entity set and $E = \{r_{ij}\} \subseteq V \times V$ represents the relations connecting the entities, the task of KG-to-text aims to generate a passage of text $\hat{Y} = (y_1, y_2, \ldots, y_n)$, that faithfully represents the information contained in a graph G. The model is given a training set $\mathcal{D} = \{(G_i, Y_i)\}$, in which the reference text Y_i may contain *hallucinated* information.

3.2 Our Framework

Standard fine-tuning approaches use a cross-entropy loss to maximize the similarity between the ground-truth text and the output text. Thus, if the ground-truth text contains hallucination, the model trained through fine-tuning also learns to generate hallucinated text. To overcome this hallucination problem, we introduce an effective fine-tuning approach that combines a contrastive loss function and a controllable text generation technique with the cross-entropy loss function. As a result, our

method can train a KG-to-text generation model to generate faithful text from a KG.

Figure 2 depicts the overall architecture of our proposed model. The following two subsections illustrate our two proposed techniques in detail.

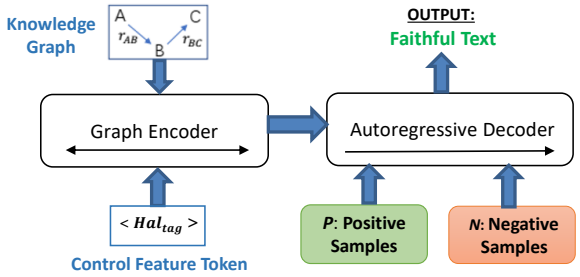


Figure 2: The overall framework of our KG-to-text model.

3.3 Minimizing Hallucinations with Contrastive Learning

Contrastive learning is a popular and effective representation learning method. Originally proposed for computer vision tasks (Khosla et al., 2020; Yang et al., 2022), contrastive learning has been successfully applied to learn representations for sentences/documents (Gao et al., 2021; Zhang et al., 2021), abstractive summarization (Liu and Liu, 2021; Cao and Wang, 2021; Wan and Bansal, 2022) and dialogue generation (Tang et al., 2022; Dziri et al., 2022; Geng et al., 2022). Inspired by them, we have utilized this learning framework to reduce hallucinations while generating text from knowledge graphs. It enables the model to differentiate between faithful information and hallucinated information in the text, which then assists the decoder in generating text that should be free of hallucinations.

For an input pair of a graph and an anchor reference text (G_i, Y_i) from the training data $\mathcal{D}$, P_i represents a set of positive samples and N_i represents a set of hallucinated summaries (i.e. negative samples). The contrastive learning objective function is formulated as follows in Equation 1:

$$L_{CL} = -\sum_{(G_i,Y_i)\in\mathcal{D}} \sum_{Y_j\in P_i} \log \frac{\exp(\cos(h_i, h_j))}{\sum_{Y_k\in N_i} \exp(\cos(h_i, h_k))} \quad (1)$$

Here, Y_j is a positive sample from the set P_i, Y_k is a negative sample from the set N_i, and h_i, h_j, h_k are the BART decoder representations of Y_i, Y_j, and Y_k respectively.

This contrastive objective function encourages the model to learn a preference for positive (faithful) summaries over negative (hallucinated) ones. While the ground-truth text in the training data $\mathcal{D}$ is noisy, it is reasonable to assume that each reference text is more faithful to the paired graph than a randomly sampled text from $\mathcal{D}$. Based on this observation, we carefully select the positive and negative samples to ensure the effectiveness of our contrastive learning technique.

Positive sample construction. Back-translation (Mallinson et al., 2017) is an effective approach for preserving meanings and providing linguistic diversity. Hence, we use NLPAug (Ma, 2019) to translate each anchor text to German and back to English and take the translated text as a positive sample for the anchor passage.

Negative sample construction. For the anchor text of a given graph, we treat the text of any other graph in $\mathcal{D}$ as a potential negative sample. We randomly select four such text to construct N for each anchor text. Dataset-specific knowledge can be easily incorporated in this approach to improve the quality of contrastive learning. For the House dataset, we adopt a simple heuristic for constructing the negative sample set. Here, we give more importance to the six major features of a house graph: (1) house location (2) house address (3) number of bedrooms (4) number of bathrooms (5) number of parking spaces, and (6) house property type. If all of these major features of a house differ from the anchor house, then the house's paired text is selected as the negative sample for the anchor house. We choose these six features as major features because information of these features is available in almost every house (91%) in the training set.

3.4 Controlling Hallucinations with Control Feature Token

In contrastive learning, we use the ground-truth reference text as a positive sample. As the ground-truth text contains hallucinations, when training with contrastive learning for generating text, the output text still contains some hallucinations. Thus, we employ a controllable text generation approach to further enhance the faithfulness of our model. Specifically, we append controllable features to the input graph in training in order to control the level of hallucination in the generated text.

Control feature token. Control feature token is a hallucination measure that quantifies how much

the given ground-truth text is faithful to the source graph. We linearized the knowledge graph (Chen et al., 2020) into a list of verbalized triples and employ BARTScore (Yuan et al., 2021) as the measure of faithfulness between the linearized graph and the corresponding ground-truth text, as it has been shown that it is closely associated with human evaluations of faithfulness (Yuan et al., 2021).

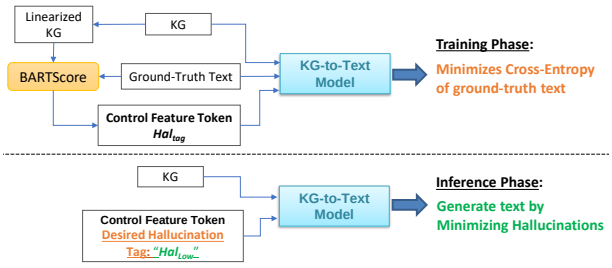


Figure 3: Controllable Text generation with Control Feature Token

Controllable generation. According to the BARTScore of the training samples, we split the samples into three buckets, where each bucket contains a list of training samples at a specific range of BARTScore. This range is chosen in a manner that ensures each bucket contains approximately an equal number of samples. These three buckets are represented using the following hallucination tags, Hal_{tag}=$\{Hal_{low}, Hal_{medium}$ and $Hal_{high}\}$ following existing work (Filippova, 2020; Zhang et al., 2022). At training time, we append the corresponding hallucination tag to the input sample according to its BARTScore. These three hallucination tags represent the three control feature tokens that act as a special input to control the level of hallucination during text generation.

Figure 3 illustrates the fine-tuning process with the control tokens. Let G and $Y = (y_1, y_2, \ldots, y_n)$ be the input sample graph and its corresponding reference text, and H be the hallucination tag (i.e. control feature token) for this input sample. Formally, we define the objective function of our fine-tuning strategy with the control token as follows:

$$L_{CE\_CtrlTok} = -\sum_{i=1}^{n} logP(y_i|y_{<i}, G, H) \quad (2)$$

Thus, during training, the model learns the mapping between the graph-text pair (G, Y) and its corresponding control token H. The model then becomes an expert at evaluating samples according to the control token. At inference time, the control token is set to the desired hallucinated value i.e., low (Hal_{low}) to generate faithful text from the KG.

The overall training objective of our proposed model is the sum of the contrastive loss and the cross-entropy loss with the control token:

$$L = L_{CL} + L_{CE\_CtrlTok} \quad (3)$$

Thus, during training, instead of blindly following the ground-truth text, the model gives more focus on the faithful parts of the text instead of the hallucinated ones. Moreover, the decoder is encouraged to generate text by minimizing hallucinations through controlled measures.

4 Experiments

4.1 Dataset

We conduct experiments and evaluation on two KG-to-text generation datasets: the House dataset (Das et al., 2021) about real-estate house listing and the GenWiki dataset (Jin et al., 2020). In both datasets, the ground-truth text contains a significant amount of hallucinated information, making the task of generating faithful text especially challenging. Thus, these datasets are the most appropriate to evaluate the performance of our proposed model. Table 1 shows the statistics of these two datasets in detail. Note that we use the "FINE" version (Jin et al., 2020) of GenWiki.

Dataset	#Relations	#KG-Text Pairs (Train / Valid / Test)
House	68	33K / 10K / 10,219
GenWiki$_{\text{FINE}}$	287	750K / 7,152 / 1,000

Table 1: Statistics of the datasets, including the total number relations and the data split

House. The dataset is prepared from the large real-estate and POI datasets of Melbourne, Australia (Das et al., 2021). It includes 53,220 records of house sales transactions from 2013 to 2015. It consists of three types of point-of-interests (POIs), namely regions, schools, and train stations, along with their corresponding features. Every sample in the dataset includes a ground-truth advertisement text describing the features of the house. However, the given ground-truth text contains a significant level of hallucinated information.

GenWiki. It is a large-scale non-parallel (Colas et al., 2021b) dataset prepared by matching Wikipedia articles with DBpedia entities (Jin et al., 2020).

House Dataset					
Model	Comparison with ground-truth text			Comparison with linearized graph	
	BLEU ↑	METEOR ↑	ROUGE-L ↑	BARTScore ↑	FactCC ↑
Ground-truth text (5K samples)	-	-	-	-4.564	48.48
JointGT (Ke et al., 2021)	**3.61**	11.96	**18.62**	-3.685	49.53
GAP (Colas et al., 2022)	3.47	**12.05**	18.16	-3.666	52.71
GMP (Han and Shareghi, 2022)	3.09	10.73	16.23	-3.941	48.47
Our Full Model	2.54	11.06	16.86	**-3.245**	**63.61**
Control token only	2.88	11.2	17.35	-3.567	52.97
Contrastive learning only	2.56	11.04	16.89	-3.247	63.04
GenWiki Dataset					
Model	Comparison with ground-truth text			Comparison with linearized graph	
	BLEU ↑	METEOR ↑	ROUGE-L ↑	BARTScore ↑	FactCC ↑
Ground-truth text (5K samples)	-	-	-	-3.464	53.80
CycleGT (Guo et al., 2020)	**41.59**	**35.72**	**63.31**	-3.276	76.86
JointGT (Ke et al., 2021)	37.93	32.60	59.06	-2.299	79.94
GMP (Han and Shareghi, 2022)	35.43	32.68	57.63	**-1.601**	76.62
Our Full Model	37.48	32.70	60.40	-2.182	**82.85**
Control token only	37.01	32.38	59.57	-2.268	81.98
Contrastive learning only	35.19	31.33	57.89	-2.309	81.48

Table 2: Results on the **House** and **GenWiki** datasets. We have used BART-base and T5-base for House dataset and Genwiki dataset respectively. **Bold** fonts denote the best results.

4.2 Baseline Models

We evaluate the performance of our proposed model against graph-to-text generation models that are based on an encoder-decoder architecture. On the House dataset, we choose three state-of-the-art models: JointGT model (Ke et al., 2021) that jointly learns the graph structure and text; GAP (Colas et al., 2022) that is aware of the graph structure; and GMP (Han and Shareghi, 2022), a self-supervised graph masking pre-training model. On the GenWiki dataset, we compare the results of the following models: the state-of-the-art unsupervised model CycleGT (Guo et al., 2020) for Genwiki dataset, JointGT (T5) model (Ke et al., 2021) and GMP (Han and Shareghi, 2022). Note that in addition to the existing state-of-the art model, GMP, we also include CycleGT as it has the best reported performance on GenWiki dataset.

4.3 Experimental Settings

We adopt JointGT (Ke et al., 2021) as our base model for fine-tuning. JointGT is initialized with the Hugging Face's pre-trained BART-base checkpoint[1] for House Dataset. For GenWiki dataset the model is initialized with the Hugging Face's pre-trained T5-base checkpoint[2]. We select the pre-trained LM BART-base or T5-base in order to do a fair comparison with the baseline models.

JointGT is pre-trained with a KGTEXT dataset (Chen et al., 2020). For contrastive learning, we use two positive samples and four negative samples for each training sample. For the House dataset, we fine-tune our model for 5 epochs; for the GenWiki dataset, we fine-tune our model for 4000 steps. The batch size is set to 32. The maximum length of linearized input graphs is 600 and the maximum length of text sequences is set to 128 tokens. We adopt Adam (Kingma and Ba, 2015) as the optimizer and set the learning rate to be 3e-5. We used one A40 48GB GPU and one A10 24GB GPU for the experiments

4.4 Main Results

We use automatic metrics to measure both fluency and faithfulness of generated text. Following existing KG-to-text work, we employ standard metrics BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004). These metrics are usually used to measure accuracy and fluency of the generated text with respect to the ground-truth text. However, as the ground-truth text contains hallucinations, we cannot verify the faithfulness of the generated text by comparing with these metrics. Thus, we use BARTScore (Yuan et al., 2021) and

[1] https://huggingface.co/facebook/bart-base

[2] https://huggingface.co/t5-base

FactCC (Kryściński et al., 2020) for comparing the generated text with the linearized input graph for measuring faithfulness. These two metrics have been widely used for measuring faithfulness in other NLP tasks (Tang et al., 2022; Gao and Wan, 2022; Cao and Wang, 2021; van der Poel et al., 2022).

The faithfulness of the reference text of the House dataset and the GenWiki dataset is also reported in Table 2, as measured by BARTScore and FactCC score. As can be seen, the reference text of both datasets contains significant amounts of hallucination (low BARTScore and FactCC scores).

Table 2 presents the results on the House and GenWiki datasets. From the results on the House dataset, we can observe that our full model achieves best results on faithfulness measures (i.e. when compared with the linearized graph), outperforming the best baseline models on BARTScore and FactCC score by 0.421 and 10.9 absolute points respectively. The performance delta on the GenWiki dataset is smaller, where our model achieves the best performance on FactCC of 1.55 points and second best performance on BARTScore. We posit the larger performance delta on the House dataset is due to it being significantly more noisy evidenced by lower BARTScore and FactCC scores.

For BLEU, METEOR and ROUGE-L, the baseline models perform modestly better than our model when comparing with the ground-truth text. This result is expected and reasonable as compared with our model, the other models tend to generate text with higher similarity with the ground-truth text, resulting in higher values as measured by these metrics. At the same time, due to the noisy nature of the reference text, a high similarity also indicates high hallucination, as discussed above.

In Section 4.5 below, we further measure the faithfulness and fluency of generated text with ChatGPT as the oracle, where we demonstrate that our model achieves superior faithfulness while maintaining fluency.

Table 3 shows a sample ground-truth text and the text generated by different models, where correct facts are highlighted in blue and hallucinated text is highlighted in red. More examples can be found in Appendix C.

4.5 ChatGPT-based Evaluation

We propose to utilize ChatGPT to further measure the factual consistency and fluency of the generated text with respect to the input graph. We randomly sample 50 houses from the House test set, and perform evaluation on the text generated by different models.

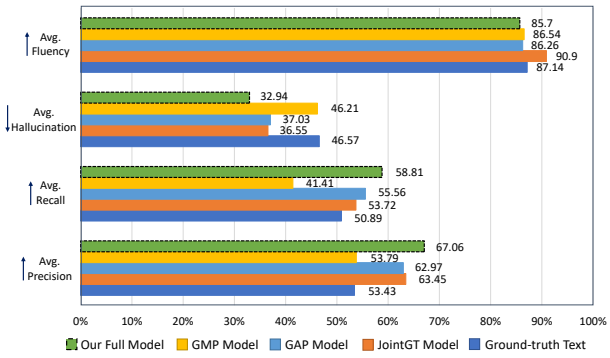


Figure 4: ChatGPT-based evaluation on 50 samples from the House test set.

To measure **fluency**, similar to (Wang et al., 2023), we prompt ChatGPT to score the fluency of the generated text. To measure **factual consistency**, we carefully design prompts to instruct ChatGPT to enumerate facts in the (linearized) graph (*# input facts*), the common facts between the graph and generated text (*# common facts*), and the hallucinated facts in the generated text (*# hallucinated facts*), respectively. By enumerating facts that are correctly generated, missing, or hallucinated, our ChatGPT-based evaluation provides better explainability of models' faithfulness. Details and examples of our prompts and ChatGPT's responses can be found in Appendix A.

In addition to enumerating the facts, ChatGPT-based evaluation provides a way to measure quantitative metrics such as precision, recall, and hallucination rates. We randomly sample 50 graph-text pairs from the test House dataset, and measure the precision (P), recall (R) and amount of hallucination (H) in generated text of these samples, which are formulated as follows: $P = \frac{\#\ common\ facts}{\#\ output\ facts}$, $R = \frac{\#\ common\ facts}{\#\ input\ facts}$, and $H = \frac{\#\ hallucinated\ facts}{\#\ output\ facts}$.

The number of output facts (*# output facts*) is computed by summing up the number of hallucinated facts (*# hallucinated facts*) and the number of common facts (*# common facts*).

Figure 4 shows the results of this analysis. It can be seen that our model outperforms all the baseline KG-to-text generation models on precision, recall and faithfulness (i.e. low hallucination) and achieves competitive scores in terms of fluency.

To determine the gap between our model and the

Ground-truth text: **Exceptional potential with elevated outlook:**Set on the high side in a wonderfully quiet and convenient cul-de-sac location, this clinker-brick home provides an ideal opportunity to renovate and extend, rebuild or redevelop (STCA) and capture the city skyline views. Immediately comfortable, the home currently offers five bedrooms (two with built-in robes) and central bathroom, two living areas and well-appointed kitchen. Other features include heating and cooling, expansive backyard, plus garage and carport via full-length side driveway. Close to Rosehill Road/Hoffmans Road buses, shops and cafes, walking distance to St Bernard???s College and Rosehill Secondary College. Currently tenanted at $1653.00pm. Land size 15.8m x 42.1m ??? 665sqm (approx). This is the one!
JointGT model: **AUCTION THIS SATURDAY AT 11:30AM!**:In one of Niddrie's most sought-after pockets, this solid-brick home presents an exciting opportunity to renovate and extend (STCA) on a deep block of land (approx 690sqm). Current floorplan includes five bedrooms (all with built-in robes) and central bathroom, formal lounge through to the kitchen/meals area, as well as a separate toilet. Other features of the home include gas heating and split-system air conditioning, plus a north-facing backyard complete with double garage (remote-controlled
GAP model: **Spacious family home in a sought-after Niddrie location!**:Set on a large block of land (approx 697sqm), this solid-brick home presents an exciting opportunity to renovate, extend or redevelop (STCA). Current floorplan includes five bedrooms (two with BIRs) and central bathroom, open-plan living/meals area and updated kitchen with stainless steel appliances. Other features include split-system heating/cooling, polished floorboards, polished hardwood floors and a double garage with internal access. Close to schools, shops and public transport, it's a must-see
GMP model: **Exceptional family living in a prime Niddrie location!**:In a quiet cul-de-sac close to schools, shops and transport, this impressive home offers an immediately comfortable family lifestyle with plenty of scope to further improve and extend (STCA). Beyond its attractive facade, the home's inviting interior comprises five bedrooms (all with built-in robes) and stylish central bathroom, spacious lounge through to the kitchen/meals area and adjoining family room. Other highlights include ducted heating and evaporative cooling, polished hardwood floors, plus a double garage (with
Our model: **SOLD by Frank Dowling Real Estate**: Located in a quiet pocket of Niddrie, this immaculately presented family home is sure to impress. Comprising of 5 good sized bedrooms with BIRs, master with WIR and walk in robe, central bathroom, open plan living and dining area, central kitchen with stainless steel appliances, family bathroom and separate toilet. Other features include gas heating, evaporative cooling, split system air conditioner, double garage with internal access and a large rear yard with rear access. Close to schools, shops and transport.

Table 3: An example of ground-truth and generated text on the House dataset. Here red colored text represents hallucinated information and blue colored text represents the faithful information.

most capable language models, we also compare our model with ChatGPT on a set of 1,000 random samples from the House dataset in different settings. A comprehensive analysis of this experiment is presented in Appendix B. As can be expected, ChatGPT achieves significantly better performance in faithfulness in zero-shot setting. However, when given noisy ground-truth text as few-shot examples, ChatGPT generates hallucinated text similar to the ground-truth text, showing that it is also prone to noise in the reference text. Our model outperforms ChatGPT in this (3-shot) setting in terms of precision and hallucination (i.e., lower hallucination).

4.6 Ablation Studies

To investigate the effect of contrastive learning and control token techniques individually, we experiment on both datasets with two configurations of our full model: one with control token only and the other one with contrastive learning only.

As we see in Table 2, both model components contribute to our model's better faithfulness, with contrastive learning making a larger impact in House dataset.

5 Conclusion

In this paper, we have proposed a novel approach to generate faithful text from a knowledge graph having noisy ground-truth text. To ensure faithful text generation, we have introduced two key ideas: (i) contrastive learning to better differentiate between faithful and hallucinated information, (ii) control token to regulate the level of hallucination in the generated text. Experimental results on two noisy KG-to-text datasets demonstrates that KG-to-text model with our framework outperforms all the baseline models in terms of faithfulness metrics. Moreover, we have proposed a novel ChatGPT based evaluation technique for an in-depth quantitative and qualitative analysis, which further verifies the superior performance of our model on

precision, recall and faithfulness.

Limitation and Future work We have applied our proposed framework only in PLM based KG-to-text encoder-decoder model. In future, we plan to explore the hallucination problem in AMR (Abstract Meaning Representations) graph datasets, which can also preserve a number of meaningful semantic relations and widely used in NLP areas.

Ethical Considerations

Our model utilizes existing pre-trained language model based KG-to-text generation model, thus the ethical concerns associated with these models would also be applicable to our proposed framework.

Acknowledgments

This material is based on research sponsored by Defense Advanced Research Projects Agency (DARPA) under agreement number HR0011-22-2-0047. The U.S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

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.

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.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. Kgpt: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2023. Toward subgraph-guided knowledge graph question generation with graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Anthony Colas, Mehrdad Alvandipour, and Daisy Zhe Wang. 2022. Gap: A graph-aware language model framework for knowledge graph-to-text generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5755–5769.

Anthony Colas, Ali Sadeghian, Yue Wang, and Daisy Zhe Wang. 2021a. Eventnarrative: A large-scale event-centric dataset for knowledge graph-to-text generation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Anthony Colas, Ali Sadeghian, Yue Wang, and Daisy Zhe Wang. 2021b. Eventnarrative: A large-scale event-centric dataset for knowledge graph-to-text generation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Sarkar Snigdha Sarathi Das, Mohammed Eunus Ali, Yuan-Fang Li, Yong-Bin Kang, and Timos Sellis. 2021. Boosting house price predictions using geo-spatial network embedding. *Data Mining and Knowledge Discovery*, 35:2221–2250.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Andrea Madotto, Osmar R Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870.

Mingqi Gao and Xiaojun Wan. 2022. Dialsummeval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Zhichao Geng, Ming Zhong, Zhangyue Yin, Xipeng Qiu, and Xuan-Jing Huang. 2022. Improving abstractive dialogue summarization with speaker-aware supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6540–6546.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

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.

Jiuzhou Han and Ehsan Shareghi. 2022. Self-supervised graph masking pre-training for graph-to-text generation. In *Empirical Methods in Natural Language Processing 2022*, pages 4845–4853. Association for Computational Linguistics (ACL).

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

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.

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.

Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2526–2538.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv e-prints*, pages arXiv–2302.

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.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893.

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.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *EMNLP 2022*. arXiv.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2022. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, pages 1–37.

Leonardo F. R. Ribeiro, Yue Zhang, Claire Gardent, and Iryna Gurevych. 2020a. Modeling global and local node contexts for text generation from knowledge graphs. *Transactions of the Association for Computational Linguistics*, 8:589–604.

Leonardo FR 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.

Leonardo FR Ribeiro, Yue Zhang, Claire Gardent, and Iryna Gurevych. 2020b. Modeling global and local node contexts for text generation from knowledge graphs. *Transactions of the Association for Computational Linguistics*, 8:589–604.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4498–4507.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *CoRR*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*.

David Wan and Mohit Bansal. 2022. Factpegasus: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Peng Wang, Junyang Lin, An Yang, Chang Zhou, Yichang Zhang, Jingren Zhou, and Hongxia Yang. 2021. Sketch and refine: Towards faithful and informative table-to-text generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4831–4843.

Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. 2022. Improving faithfulness by augmenting negative summaries from fake documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11913–11921.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086.

Xinnuo Xu, Ondřej Dušek, Verena Rieser, and Ioannis Konstas. 2021. Agggen: Ordering and aggregating while generating. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1419–1434.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen Mckeown, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430.

Haopeng Zhang, Semih Yavuz, Wojciech Kryściński, Kazuma Hashimoto, and Yingbo Zhou. 2022. Improving the faithfulness of abstractive summarization via entity coverage control. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 528–535.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

A Prompt Design for ChatGPT-based Evaluation

The prompt templates are shown in Figure 5.

Listing the facts of a graph: Here we give ChatGPT an input linearized graph and ask it to "list the features one by one from the INPUT" (Figure 5-Left). Figure 6 shows an example of this prompt to ChatGPT and its response for a sample from the House test set. ChatGPT has made no error in all 50 test samples of House data.

Listing the common facts: ChatGPT was unable to correctly list the common facts between the linearized input graph and the generated text. Hence, we prompt ChatGPT for each fact listed in the input, whether that fact is included in the output. Here, each fact (or "feature") represents a single triple of the input linearized graph (Figure 5-Middle). Then, we count the answer with a "yes" response from ChatGPT. On average, ChatGPT makes 2-3 mistakes per sample. Figure 7 shows an example of this prompt and and ChatGPT's response. The red colored text indicates the mistakes done by ChatGPT.

Listing the hallucinated facts: Here, we prompt ChatGPT to list both the extrinsic and intrinsic hallucination facts in the generated text by providing ChatGPT with an input (linearized graph) and an output (generated text). Firstly, to list the extrinsic hallucination facts we instruct ChatGPT to "List the features one by one from the OUTPUT that is not mentioned in the INPUT". Secondly, to list the intrinsic hallucination facts we instruct ChatGPT to "List the features one by one from the OUTPUT that is contradictory to the INPUT" (Figure 5-Right). Here, ChatGPT makes no mistakes in the 50 House test samples. Figure 8 illustrates the steps with an example and ChatGPT's response.

B Comparing Our Result with ChatGPT

We randomly take 1000 sample graphs from the House dataset. Our experiments are conducted using the API of Chat- GPT (gpt-3.5-turbo) model. We input ChatGPT the sample graphs in a linearized format and asked to summarize the linearized graphs in a real-estate advertising format. We experiment with ChatGPT-ZeroShot (without giving any reference text) , ChatGPT-k-FewShot, (where k represents the number of noisy ground-truth text sample is given to ChatGPT as a reference in addition to the input linearized graph) and compare these with our full model.

Table 4 shows that in terms of faithfulness metrics (BARTScore), ChatGPT-ZeroShot has the best performance. This is because, ChatGPT is a large model and ChatGPT-ZeroShot generates text without taking any noisy ground-truth text as a reference. Whereas, our model is a small (BART-base/T5-base) language model and the model is trained with the full noisy training House dataset. We also notice that the performance of ChatGPT-k-FewShot drops with the increase of number of noisy reference text samples. Thus, the more we increase the number of noisy ground-truth texts as a reference to ChatGPT, the more ChatGPT generates hallucinated text similar to ground-truth text. That's why the BLEU, METEOR and ROUGE-L scores increase and BARTscore, FactCC scores decrease with the increase of few shot samples.

We also compare the results using ChatGPT-based evaluation. Table 5 shows the average of precision, recall and hallucinations which we compute using ChatGPT. The results also show that ChatGPT-ZeroShot performs best in all metrics as usual. Our model outperforms ChatGPT-3-FewShot in terms of precision (higher precision) and hallucination (lower hallucination).

Performance Based on Salient Facts: We rank in descending order the features (type-wise) of the house graph based on their frequency of occurrence in the House trainining dataset. We take top ten features as *salient* facts. The salient facts are: 1) house_location, 2) house_property-type, 3) num. of bedrooms, 4) num. of bathrooms, 5) num of parking spaces, 6) has_ac, 7) has_dining, 8) has_heating, 9) has garage_spaces and 10) nearest_train_station. Using ChatGPT, we enumerate the presence of these facts and measure salient precision, $P_{salient}$ and salient recall, $R_{salient}$ as follows.

$$P_{salient} = \frac{\textit{\# salient common facts}}{\textit{\# output facts}} \tag{4}$$

$$R_{salient} = \frac{\textit{\# salient common facts}}{\textit{\# salient input facts}} \tag{5}$$

The results from Table: 6 shows that our model achieves the best average salient precision, $P_{salient}$, and ChatGPT-ZeroShot achieves the best average salient recall. The reason behind this result is that ChatGPT-ZeroShot generated output text contains mostly all the facts from the input graph, whereas

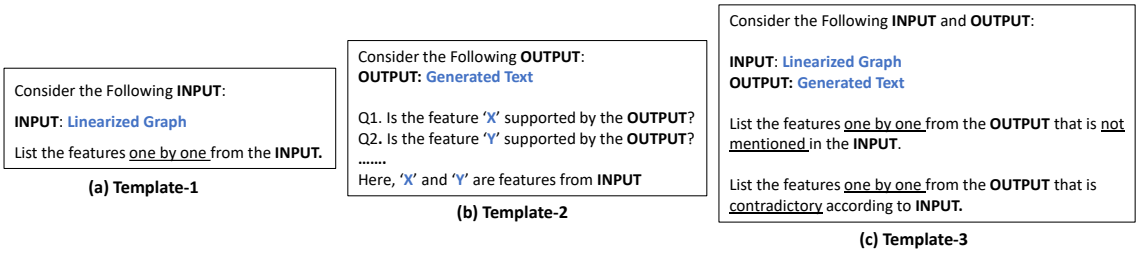


Figure 5: Prompt templates for enumerating facts using ChatGPT. Template-1 (left) is to enumerate facts in the input (linearized graph). Template-2 (middle) is to enumerate common facts between the input (linearized graph) and the output (generated text). Template-3 (right) is to enumerate hallucinated facts in the output (generated text).

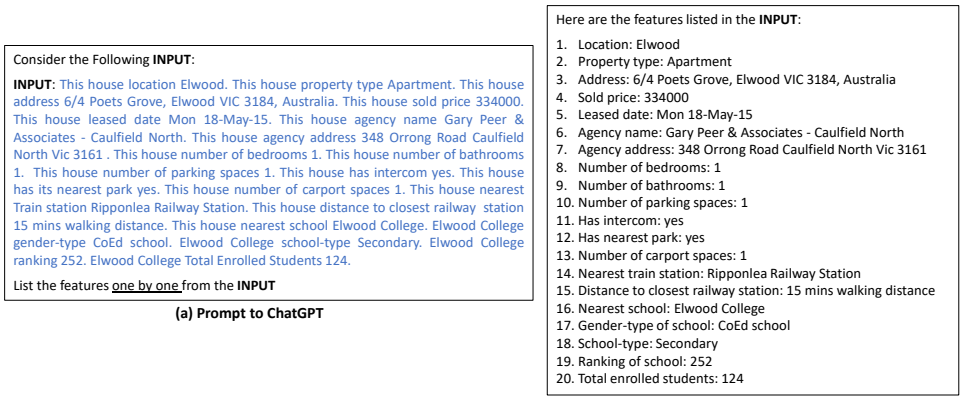
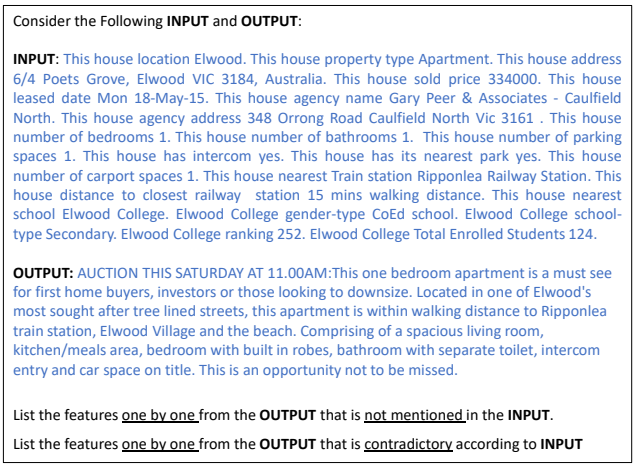
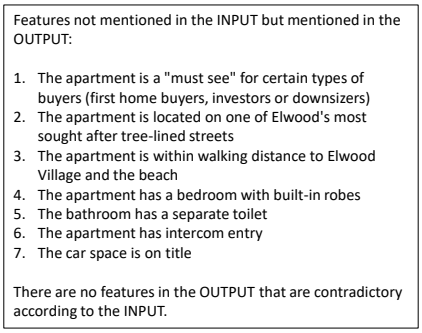


Figure 6: An Example of Template-1

Figure 7: An Example of Template-2



(a) Prompt to ChatGPT



(b) Response from ChatGPT

Figure 8: An Example of Template-3

Generation Model	Comparison with ground-truth text			Comparison with linearized graph	
	BLEU ↑	METEOR ↑	ROUGE-L ↑	BARTScore ↑	FactCC ↑
ChatGPT-ZeroShot	1.21	11.86	12.91	**-2.389**	71.02
ChatGPT-1-Shot	1.95	12.73	15.02	-2.872	**76.34**
ChatGPT-2-Shot	2.06	12.67	15.58	-2.937	72.02
ChatGPT-3-Shot	2.25	**13.31**	15.76	-3.036	73.88
Our Full Model	**2.68**	11.21	**17.10**	-3.246	62.84

Table 4: Results on 1000 test samples from the House dataset. **Bold** fonts denote the best results.

Generation Model	Avg. Precision	Avg. Recall	Avg. Hallucination
ChatGPT-ZeroShot	**73.28**	**88.21**	**26.71**
ChatGPT-3-Shot	65.45	64.39	34.55
Our Full Model	67.06	58.81	32.94

Table 5: ChatGPT Evaluation Results based on 50 samples from the House Dataset. **Bold** fonts denote the best results.

our model generated output text gives more focus on the salient facts.

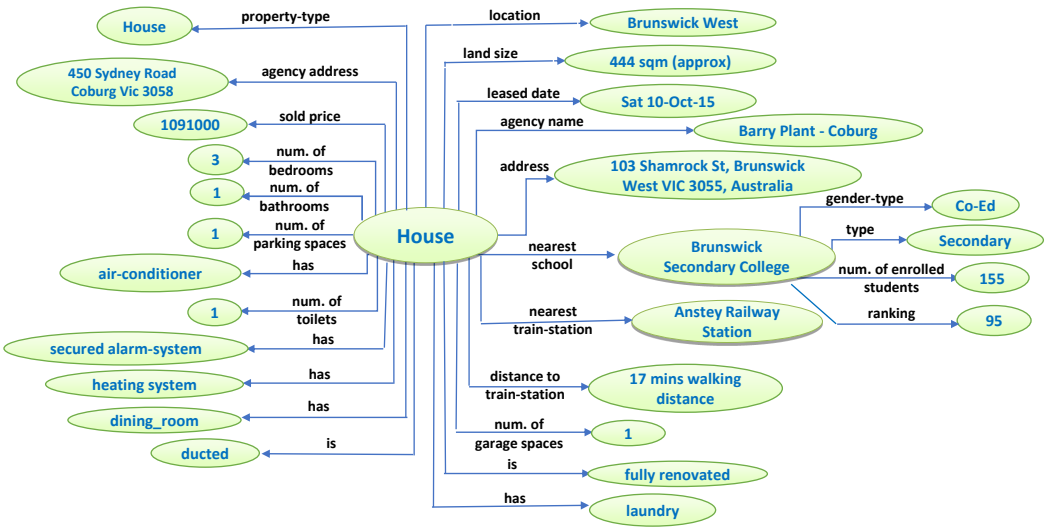
C Generated Samples

Figure 9 and Figure 10 show qualitative examples of sample graphs, the ground-truth texts and the texts generated by different models on House dataset and Genwiki dataset, respectively.

Generation Model	Avg. Salient Precision	Avg. Salient Recall
ChatGPT-ZeroShot	26.75	**92.66**
ChatGPT-3-FewShot	30.27	86.36
Our Full Model	**31.64**	77.16

Table 6: ChatGPT Evaluation Results based on 50 samples from the House dataset considering salient features. **Bold** fonts denote the best results.

House Knowledge Graph:



Ground-truth text:

AUCTION THIS SATURDAY 10TH OCTOBER AT 1PM: Superbly blending charm with a modern sensibility and fresh infusion of light, this fully renovated three bedroom Californian Bungalow offers stylish family living in a sought after locale.An expansive open plan lounge and dining room provides spacious living while showcasing a marvellous kitchen boasting 700mm stainless steel appliances, abundant storage and feature island bench, ensuring effortless entertaining.A serene second living room, boasting leafy garden views offers retreat from the bustle of family life.Three spacious bedrooms are serviced by a modern bathroom, which cleverly incorporates an oversized European laundry.Brilliantly positioned to transport, schools, cafes, vibrant Sydney Road, CityLink and featuring ducted heating, alarm, split system air-conditioning, BIR's, ceiling fans and carport with OSP, stylish family living starts here.

JointGT model:

AUCTION THIS SATURDAY AT 11:00AM: This renovated three bedroom brick veneer home is sure to appeal to first home buyers, downsizers and investors alike.The home comprises of three double bedrooms (all with built-in robes), spacious lounge and dining room, renovated kitchen with stainless steel appliances, separate laundry, ducted heating, evaporative cooling, alarm system and a single garage with internal access.The property also features a large rear garden with established gardens and a bungalow.

GAP model:

Renovated And Ready To Move In: This beautifully renovated brick veneer home is ready to move in and enjoy. Featuring a spacious living/dining area, renovated kitchen with stainless steel appliances and adjoining meals area, three double bedrooms with built-in robes, central bathroom and separate laundry. Other features include ducted heating, evaporative cooling, alarm system, double garage with internal access, polished floorboards, polished hardwood floors, security shutters and much more. Close to Sydney Road shops, schools, public transport and easy access to CBD.

GMP model:

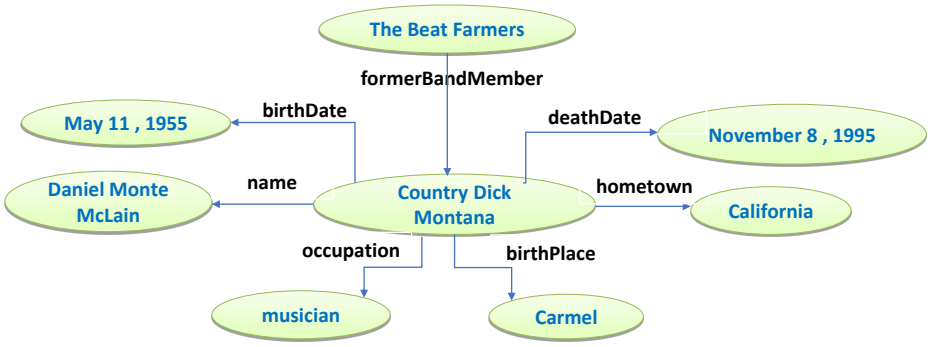
SOLD BY HARRY KONTOSSIS: Sale by SET DATE 10/10/2015 (unless sold prior) Situated in a quiet cul-de-sac in the heart of Brunswick West, this immaculately presented three bedroom, two bathroom home is sure to appeal to first home buyers, investors and downsizers alike.Accommodation includes a formal lounge and dining room, well-appointed kitchen with stainless steel appliances, adjoining family/meals area, three double bedrooms with built-in robes, central bathroom, separate toilet.

Our framework:

Renovated Family Home in the Heart of Brunswick West: This beautifully renovated family home is set on a generous allotment of 444 sqm (approx) and is ideal for first home buyers, downsizers or investors looking for a low maintenance lifestyle. Comprising of 3 spacious bedrooms (master with WIR and walk in robe), formal lounge and dining room, updated kitchen with stainless steel appliances, central bathroom and laundry. Other features include ducted heating, evaporative cooling, split system air conditioner, ducted cooling, alarm system, double garage and much more.

Figure 9: Example of generated text on the House dataset. Here red colored text represents hallucinated information and blue colored text represents the faithful information

Genwiki Knowledge Graph:



Ground-truth text:

Daniel Monte McLain (May 11 , 1955 – November 8 , 1995) , known by the stage name Country Dick Montana , was a musician best known as a member of The Beat Farmers . Montana was born in Carmel , California .

JointGT model:

Montana was born on May 11 , 1955 in Carmel , California .

CycleGT model:

Daniel Monte McLain (May 11 , 1955 in Carmel , Montana – November 8 , 1995 in Carmel , California) was a musician , best known as the founder of the band Country Dick Montana .

GMP model:

Daniel Monte McLain (May 11, 1955 – November 8, 1995), known professionally as Country Dick Montana, was an American singer, songwriter, and musician.

Our framework:

Daniel Monte McLain (May 11 , 1955 – November 8 , 1995) was an American musician .

Figure 10: Example of generated text on the Genwiki dataset. Here red colored text represents hallucinated information and blue colored text represents faithful information.