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

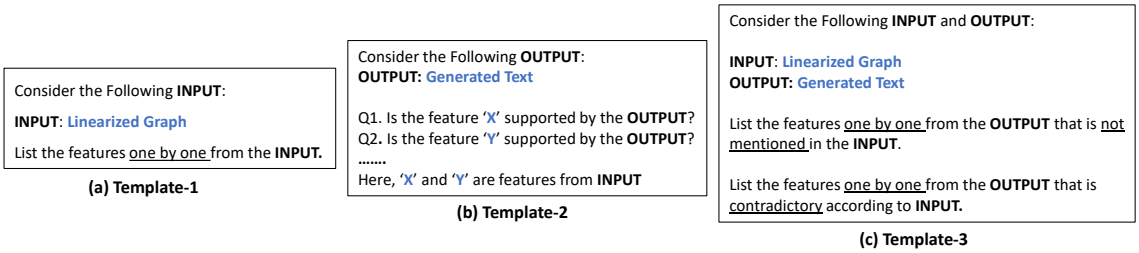


Figure 5: Prompt templates for enumerating facts using ChatGPT. Template-1 (left) is to enumerate facts in the input (linearized graph). Template-2 (middle) is to enumerate common facts between the input (linearized graph) and the output (generated text). Template-3 (right) is to enumerate hallucinated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

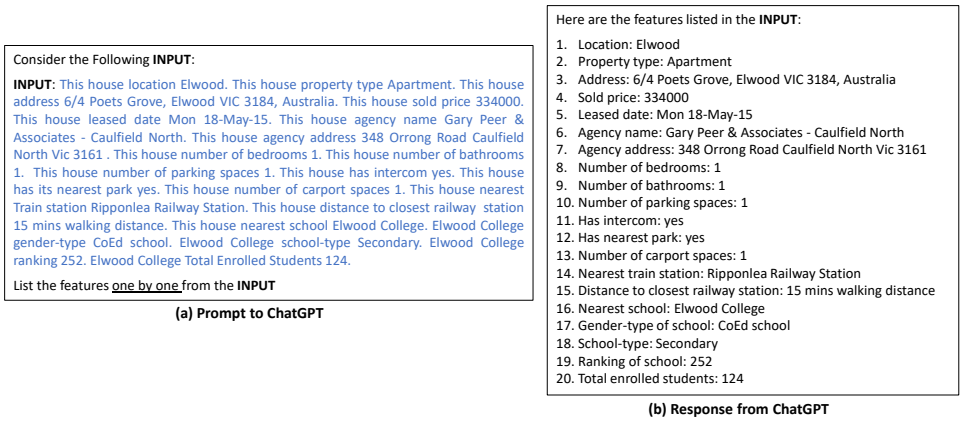


Figure 6: An Example of Template-1

Figure 7: An Example of Template-2

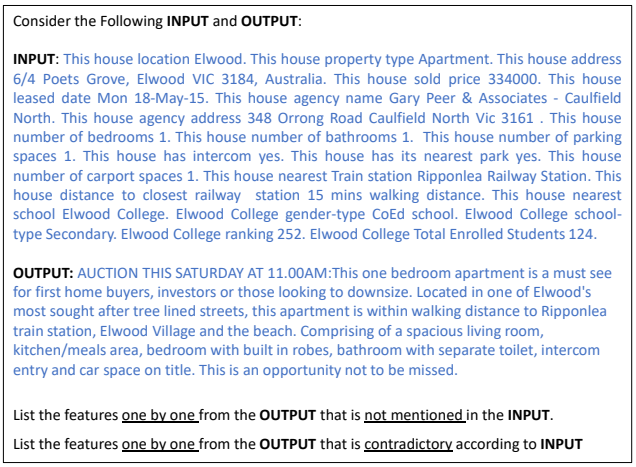


Figure 8: An Example of Template-3

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.

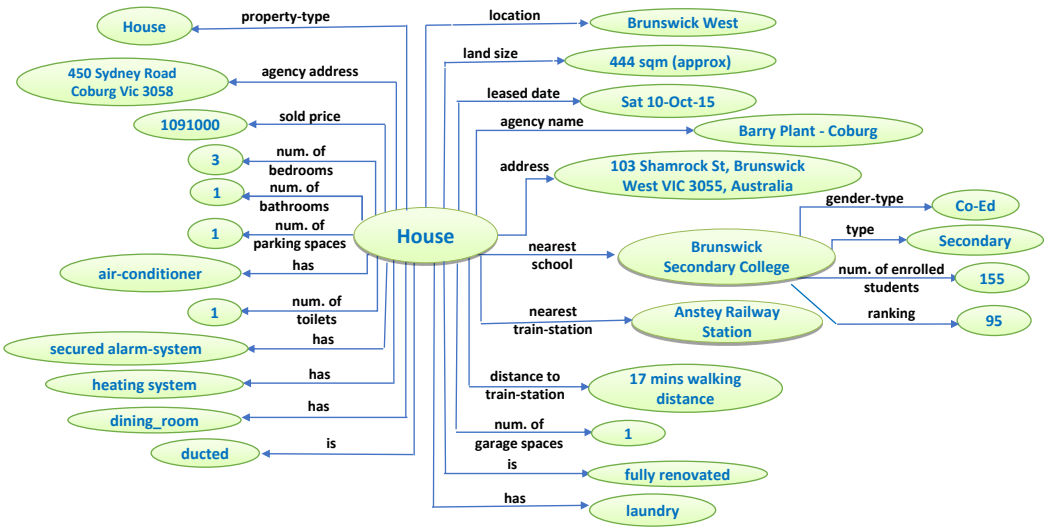
C Generated Samples

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:

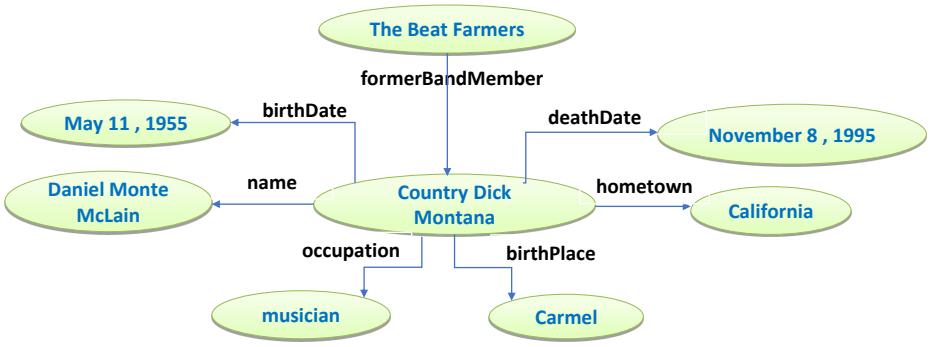
SOLD BY HARRY KONTOSSIS: Sale by SET DATE 10/10/2015 (unless sold prior) Situated in a quiet cul-de-sac in the heart of Brunswick West, this immaculately presented three bedroom, two bathroom home is sure to appeal to first home buyers, investors and downsizers alike.Accommodation includes a formal lounge 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.