A Comparative Analysis of Conversational Large Language Models in Knowledge-Based Text Generation

Anonymous ACL submission

Abstract

Generating natural language text from graph-structured data is essential for conversational information seeking. Semantic triples derived from knowledge graphs can serve as a valuable source for grounding responses from conversational agents by providing a factual basis for the information they communicate. This is especially relevant in the context of large language models, which offer great potential for conversational interaction but are prone to hallucinating, omitting, or producing conflicting information. In this study, we conduct an empirical analysis of conversational large language models in generating human-readable text from semantic triples. We compare four large language models of varying sizes with different prompting techniques. Through a series of benchmark experiments, we analyze the models' performance and identify the most common issues in the generated predictions. Our findings demonstrate that the capabilities of large language models in triple verbalization can be significantly improved through few-shot prompting, efficient fine-tuning, and post-processing techniques, particularly for smaller models that exhibit lower zero-shot performance.

1 Introduction

Accessing structured information through natural language interfaces has garnered significant research interest in natural language processing (NLP) (Radlinski and Craswell, 2017; Aliannejadi et al., 2021). These search-oriented conversational interfaces are often connected to structured data sources like knowledge graphs. However, a key challenge lies in mediating between natural language, in which users express their queries, and machine-readable knowledge representations. The task of data-to-text generation focuses on this issue, taking structured data as input to produce coherent, human-readable text, which has been extensively studied with approaches ranging from rule-based to supervised neural network-based techniques.

Over the last years, the field of NLP has witnessed a shift in methodologies with the advent of pre-trained large language models (LLMs). Unlike traditional supervised learning approaches that rely on annotated datasets, LLMs are trained in a self-supervised manner, predicting tokens within vast amounts of unlabeled data. Combined with scaling up the model size and training corpora, this approach has demonstrated remarkable emergent capabilities of LLMs and their prowess in multi-task learning (Radford et al., 2019; Brown et al., 2020). An advantage of LLMs lies in prompt-based (in-context) learning. Through carefully defined prompts, these foundation models can perform multiple tasks like question-answering or text summarization (Liu et al., 2023). More recently, there has been a growing interest in optimizing LLMs for conversational interactions by pre-training on dialogue corpora, instruction fine-tuning, and reinforcement learning from human feedback (Thoppilan et al., 2022; OpenAI, 2022). Although LLMs offer tremendous potential for conversational interaction, owing to their ability to produce responses for arbitrary input texts, they have known limitations, such as the risk of hallucinating or omitting important information and a lack of transparency regarding the origins of information sources from which they derive their outputs (Dou et al., 2022; Ji et al., 2023). In order to mitigate these limitations, it becomes imperative to ground their generated outputs in verifiable factual data from knowledge graphs. However, there has been insufficient systematic investigation into their proficiency in verbalizing graph-structured data input.

To assess LLMs in knowledge-based text generation, we compare four models of different sizes and training objectives, with a primary focus on models optimized for conversational interaction. Based on the popular WebNLG benchmark dataset, we evaluate the models' performance in generating natural language text from semantic triples. Through multiple experiments, we analyze different configurations of models and prompting techniques, discussing insights about their individual capabilities and limitations. Our contributions include: (1) creating a benchmark to evaluate LLMs on the WebNLG dataset, (2) comparing model performance through automatic reference-based metrics and human evaluation, and (3) providing insights on their reliability in triple-to-text generation. To ensure reproducibility, we publish our source code and datasets in an anonymous GitHub repository.[1]

2 Related Work

Existing works from the NLP literature have explored knowledge-based text generation, with significant advancements driven by new deep learning architectures and fine-tuning language models on downstream tasks (Li et al., 2021). For triple-to-text generation, many evaluations use the established WebNLG benchmark (Colin et al., 2016). Several studies have focused on comparing neural pipeline versus end-to-end approaches, assessing supervised versus unsupervised training regimes and developing frameworks for making text generation more controllable (Castro Ferreira et al., 2019; Schmitt et al., 2020; Su et al., 2021).

Concerning pre-trained language models, Chen et al. (2020) were among the first to propose the task of few-shot natural language generation. With just 200 table-to-text training examples, their approach achieves strong performance and good generalization. By collecting a novel dataset and experimenting with few-shot fine-tuning, Kasner et al. (2023) demonstrate that pre-trained language models trained with a diverse set of labels exhibit robustness in verbalizing knowledge graph relations, being capable of generalizing to novel domains. Similar to our work, Han et al. (2023) assess the capabilities LLMs but for text-to-graph generation with the model ChatGPT. They develop a prompting framework with iterative verification, improving the quality of generated outputs. In contrast, our objective is to achieve a comprehensive understanding of conversational LLMs for triple verbalization rather than solely concentrating on individual use cases or models. To the best of our knowledge, we are the first to conduct a comparative analysis of conversational LLMs and prompt configurations on the task of triple-to-text generation.

[1]GitHub: https://github.com/CS-Lab-Study/LLM-D2T

3 Experiments

Experimental Setup We conduct our experiments on the **WebNLG+ 2020** dataset, a DBpedia-based triple-to-text benchmark with 1,779 test examples (Castro Ferreira et al., 2020). As evaluation metrics, we calculate the lexical similarity between model outputs and human annotations using **BLEU** (Papineni et al., 2002), **METEOR** (Banerjee and Lavie, 2005), and **TER** (Snover et al., 2006). Since these metrics mainly focus on lexical overlaps, we also use the **BERTScore-F1** metric, which captures semantic similarity (Zhang et al., 2020).

As a commercial state-of-the-art LLM, we include **GPT-3.5-Turbo** (**ChatGPT**) (OpenAI, 2022) in our comparison. It is optimized for conversations and has demonstrated remarkable zero-shot performance on various NLP tasks. We ran our experiments with the latest model released in June 2023 (GPT-3.5-Turbo-0613). Further, we opted to test **LLaMA**, a collection of open-source LLMs from Meta (Touvron et al., 2023), achieving competitive performance on benchmarks. We include three model variations with 7B parameters of the first LLaMA version. In addition to the non-conversational base model, we tested a fine-tuned model which was trained on 26,422 WebNLG examples in chat completion format. The training was done through **low-rank adaptation (LoRA)**, a method that fine-tunes only a subset of the model's parameters, referred to as low-rank matrices, rather than updating the entire parameter space, improving the fine-tuning efficiency (Hu et al., 2022). Another fine-tuned LLaMA model we included is **Vicuna**. It was trained on a corpus of around 70K user-shared ChatGPT conversations crawled from the ShareGPT website (Chiang et al., 2023).

The LLaMA and Vicuna models are prompted in the chat completion structure of the FastChat[2] platform, replicating OpenAI's chat completion API endpoint with a structured list of system, user, and assistant messages. We set the token limit to 128 and the temperature parameter to 0, maximizing deterministic generation by favoring high-probability words. The zero-shot prompt contains only a system message with a triple verbalization instruction. The few-shot prompt expands the instruction with three in-context examples provided as user and assistant messages. Table 2 in Appendix A displays each prompt in full length.

[2]FastChat: https://github.com/lm-sys/FastChat

Model	Zero-Shot Prompt				Few-Shot Prompt			
	BLEU	METEOR	TER	BERTScore	BLEU	METEOR	TER	BERTScore
LLaMA-7B	0.06	0.21	1.03	0.84	0.11	0.26	1.03	0.85
LLaMA-7B + PP	0.15	0.25	0.76	0.89	0.38	0.36	0.53	0.94
Vicuna-7B	0.27	0.35	0.68	0.92	0.39	0.38	0.64	0.93
Vicuna-7B + PP	0.27	0.35	0.68	0.92	0.43	0.39	0.51	0.95
GPT-3.5-Turbo	0.41	**0.41**	0.56	0.95	0.39	0.40	0.65	0.94
GPT-3.5-Turbo + PP	0.41	**0.41**	0.56	0.95	0.44	**0.41**	0.50	0.95
LoRA-7B	0.47	0.40	0.55	0.94	0.47	0.40	0.55	0.94
LoRA-7B + PP	**0.52**	**0.41**	**0.42**	**0.96**	**0.53**	**0.41**	**0.42**	**0.96**
Copy-Baseline	0.02	0.02	0.95	0.79	0.02	0.02	0.95	0.79

Table 1: Zero-shot and few-shot performance metrics on WebNLG test set evaluated by BLEU, METEOR, TER, and BERTScore-F1 (+ PP denotes post-processed model output). Bold values indicate the best value per metric.

Results of Performance Metrics Table 1 summarizes the calculated metrics. The Copy-Baseline denotes copying the triples as output without modifications. We distinguish between scores for raw and post-processed (+ PP) outputs. Post-processing involved removing in-context examples or instruction parts from the input prompt which were repeated by some models in the generated output.

Examining the scores, LoRA-7B demonstrates superior performance compared to the other models. Even without few-shot examples, it effectively learned from fine-tuning to handle the triple verbalization task, gaining only a minor performance increase through few-shot prompting. The second ranking model GPT-3.5-Turbo shows similar scores, which is remarkable because it was not explicitly trained for triple-to-text generation. Notably, Vicuna achieves a performance level almost on par with the much bigger GPT-3.5-Turbo model when it was provided with in-context examples and the output was post-processed. In the zero-shot setting, Vicuna could not match the scores of GPT-3.5-Turbo but outperformed LLaMA-7B. Although LLaMA is the worst-performing model, it claims the most significant improvements through few-shot prompting and post-processing, with scores not too far from Vicuna. The metrics collectively suggest that all tested LLMs can generate reasonable output text from knowledge graph triples. Besides, we observe that while all models show improvements with few-shot prompting or post-processing, models trained on conversations like Vicuna require less post-processing and exhibit better zero-shot proficiency, resulting in comparatively smaller performance gains from post-processed outputs or in-context examples.

Analysis and Discussion The WebNLG triple verbalization task involves different subtasks, such as segmentation of the input data, lexicalization of the DBpedia properties, information aggregation, and surface realization of grammatically correct text (Colin et al., 2016). All of these subtasks are handled by LLMs in an end-to-end manner. In direct comparison to state-of-the-art models evaluated on WebNLG like **Control Prefixes** (BLEU: 0.62, METEOR: 0.45, TER: 0.35) from Clive et al. (2022) or **T5-Large+Wiki+Position** (BLEU: 0.61, METEOR: 0.44, TER: 0.36, BERTScore: 0.96) from Wang et al. (2021), the LLMs' lexical similarity metrics are worse. Yet, when looking at semantic similarity, the BERTScore metric of the LoRA-7B model is identical with 0.96. We hypothesize that the lower lexical similarity is partly caused by the concise writing style of the WebNLG human ground-truth verbalizations, aggregating as much information as possible in succinct sentences. While many WebNLG annotations are as short as possible (e.g., *"The 98.0 minute film Super Capers starring Danielle Harris was written by the director Ray Griggs."*), the more verbose output of LLMs like GPT-3.5-Turbo consists of multiple sentences (e.g., *"Danielle Harris stars in the movie Super Capers. The writer of the movie is Ray Griggs. The movie has a runtime of 98.0 minutes."*). This concise writing style can be better learned and replicated by LoRA and other fine-tuned models.

With a larger number of input triples, models struggle more to transform structured information into cohesive text. Figure 1 illustrates the decreasing model performance when confronted with multiple triples. While all four LLMs follow the same trend, the performance loss seems to be a tapering decrease. Since aggregating information into short sentences is also desired in conversational user interactions, we compared the sentence count of generated predictions for each model regarding the number of input triples. As can be discerned from Figure 2 in Appendix A, the fine-tuned

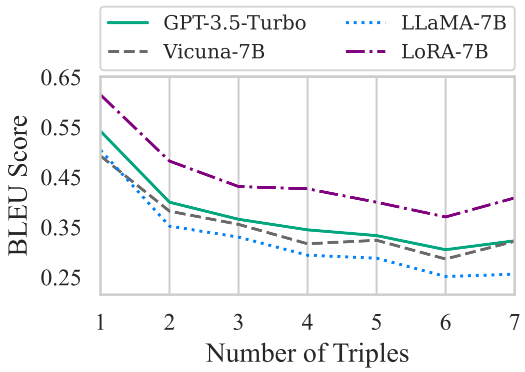

Figure 1: Comparison of BLEU score by number of triples for few-shot models with post-processing.

LoRA model produces sentences in direct proportion to the number of input triples in alignment with the human annotations. Vicuna and GPT-3.5-Turbo, which have been explicitly trained on conversations, exhibit a similar generation behavior. While LoRA produces the fewest sentences, Vicuna seems to be a bit less verbose than GPT-3.5-Turbo. In contrast, text outputs from LLaMA contain, on average, the largest number of sentences and show a much higher variance. This suggests that fine-tuning LLMs on instructions from dialogue corpora improves adherence to concise triple verbalization.

After conducting the automatic evaluation, we manually examined the model predictions to gauge their reliability and grouped the most common issues into five types as presented in Table 4 in Appendix A. For example, the LLMs sometimes misinterpreted the prompt, failed to lexicalize triples correctly, or produced inaccurate information. Most of these issues occurred in zero-shot predictions from LLaMA or Vicuna, whereas GPT-3.5-Turbo produced the most reliable outputs. To obtain deeper insights into the model-specific occurrence rates of the issue types, two researchers jointly evaluated a sample of 75 zero- and 75 few-shot predictions for the lowest averaged BLEU and METEOR scores across all models. Looking at Table 3, it can be seen that LLaMA has the highest incidence of issues from all types, followed by Vicuna and then LoRA with better reliability, and GPT-3.5-Turbo as the most dependable model.

As to be expected from instruction-tuned and fine-tuned models, LoRA, Vicuna, and GPT-3.5-Turbo demonstrate greater ability in generating zero-shot output that aligns with the given prompt. Conversely, LLaMA tended to misinterpret the prompt, failing to produce the desired output format in nearly two-thirds of the evaluated instances (0.65). Interestingly, off-prompt issues could be effectively addressed in all models by including few-shot examples in the prompt. While few-shot prompting reduced off-prompt generations and caused the LLMs to produce actual sentences based on the graph triples, this led to a relative increase of inaccurate generations, such as hallucinated information, twisted numbers, or often omitting facts from the input triples. Occasionally, the relationships within these triples were also compromised. The rate of inaccurate zero-shot output in LLaMA (0.60) and Vicuna (0.41) was three to four times higher in comparison to GPT-3.5-Turbo (0.13).

Another issue type where the usefulness of few-shot examples became evident is unlexicalized triples, meaning the translation of entities and relations into their intact word form. This was observed across all models except LoRA, with LLaMA and Vicuna particularly affected. Providing in-context examples with lexicalized triples could completely resolve unlexicalized triples for all models. Problems with redundancy, which involves the unnecessary repetition of information, are mostly associated with LLaMA. This was due to some instances where LLaMA became stuck in a loop, repeatedly generating the same sequence until the maximum token limit was reached. In contrast, this issue type appears to be less of a problem for the other models. Lastly, there are rare cases in which the LLM generated output in a language other than the prompt language English. This happened, for example, when most of the input triples contained words in Spanish. Only Vicuna faced translation issues in our benchmark test, specifically in zero-shot scenarios. This behavior may be attributed to its fine-tuning dataset with translation instructions.

4 Conclusion

We compared the abilities of LLMs in triple-to-text generation. Our findings indicate that even smaller 7B-LLMs exhibit reasonable performance in verbalizing triples, conveying the intended meanings and facts in a sensible manner, although they might not always be factually accurate or perfectly replicate the writing style of human references. We also discussed model-specific differences and common generation errors that can be mitigated through few-shot prompting and post-processing. In future work, we plan to investigate how our findings generalize to more complex graph data structures.

5 Limitations

Our comparative analysis has certain limitations. We focus solely on text generation based on knowledge graph triples, and we acknowledge that verbalizing entire subgraphs or producing graph queries are other important tasks worth exploring. Nonetheless, by studying semantic triples, we can still derive valuable insights about the performance of LLMs for processing more complex graph data structures. In that regard, it is recommended to expand the comparison with human evaluations that go beyond automatically calculated metrics and to assess more models, particularly those trained on source code or documents with structured data.

Further, the employed test dataset is limited to English triples. Since pre-training corpora of LLMs primarily consist of English text data, they likely work better where entities and relations correspond to meaningful English words or morphemes. Consequently, it is to be expected that LLMs exhibit worse performance on multilingual benchmarks with more morphologically rich languages, such as Russian, which is also part of the WebNLG dataset.

6 Ethical Considerations

Our experiments were conducted on the publicly available WebNLG dataset, ensuring that no demographic or identifying information about individuals was processed or disclosed. Because our focus was not on addressing well-documented issues like privacy or biases associated with LLMs, we acknowledge potential risks and concerns in line with similar studies dealing with LLMs. The experiments with LLaMA, LoRA, and Vicuna were executed on a single NVIDIA V100 GPU and required relatively low computational resources, with around one GPU hour of inference time per model.

References

Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas, and Nick Craswell. 2021. Analysing mixed initiatives and search strategies during conversational search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 16–26, New York, NY, USA. Association for Computing Machinery.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. 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 552–562, Hong Kong, China. Association for Computational Linguistics.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020. Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Online. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *LMSYS Org Blog*.

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 363–382, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Emilie Colin, Claire Gardent, Yassine M'rabet, Shashi Narayan, and Laura Perez-Beltrachini. 2016. The WebNLG challenge: Generating text from DBPedia data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167, Edinburgh, UK. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Jiuzhou Han, Nigel Collier, Wray Buntine, and Ehsan Shareghi. 2023. Pive: Prompting with iterative verification improving graph-based generative capability of llms. *arXiv:2305.12392*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Zdeněk Kasner, Ioannis Konstas, and Ondrej Dusek. 2023. Mind the labels: Describing relations in knowledge graphs with pretrained models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2398–2415, Dubrovnik, Croatia. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji rong Wen. 2021. Pretrained language models for text generation: A survey. In *International Joint Conference on Artificial Intelligence*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 117–126, New York, NY, USA. Association for Computing Machinery.

Martin Schmitt, Sahand Sharifzadeh, Volker Tresp, and Hinrich Schütze. 2020. An unsupervised joint system for text generation from knowledge graphs and semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7117–7130, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv:2201.08239*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

Qingyun Wang, Semih Yavuz, Xi Victoria Lin, Heng Ji, and Nazneen Rajani. 2021. Stage-wise fine-tuning for graph-to-text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 16–22, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

The Appendix provides further insights into the results of our research, including the model prompts in full length (Table 2), an overview of common issue types identified in the predictions along with their relative frequency (Table 3 and 4), and the distribution of generated sentences per model (Figure 2).

Prompt Type	**Prompt Content**
Zero-shot	`SYSTEM`: Generate a concise text for the given set of triples. Ensure that the generated output only includes the provided information from the triples.
Few-shot	`SYSTEM`: Generate a concise text for the given set of triples. Ensure that the generated output only includes the provided information from the triples. `USER`: Input triples: [{'object': 'Mike_Mularkey','property': 'coach','subject': 'Tennessee_Titans'}] `ASSISTANT`: Output text: Mike Mularkey is the coach of the Tennessee Titans. `USER`: Input triples: [{'object': 'Albert_E._Austin', 'property': 'successor', 'subject': 'Alfred_N._Phillips'}, {'object': 'Connecticut', 'property': 'birthPlace', 'subject': 'Alfred_N._Phillips'}, {'object': 'United_States_House_of_Representatives', 'property': 'office', 'subject': 'Alfred_N._Phillips'}] `ASSISTANT`: Output text: Albert E. Austin succeeded Alfred N. Phillips who was born in Connecticut and worked at the United States House of Representatives. `USER`: Input triples: [{'object': 'College_of_William_&_Mary', 'property': 'owner', 'subject': 'Alan_B._Miller_Hall'}, {'object': '2009-06-01', 'property': 'completionDate', 'subject': 'Alan_B._Miller_Hall'}, {'object': '101 Ukrop Way', 'property': 'address', 'subject': 'Alan_B._Miller_Hall'}, {'object': 'Williamsburg,_Virginia', 'property': 'location', 'subject': 'Alan_B._Miller_Hall'}, {'object': 'Robert_A._M._Stern', 'property': 'architect', 'subject': 'Alan_B._Miller_Hall'}] `ASSISTANT`: Output text: The Alan B Miller Hall's location is 101 Ukrop Way, Williamsburg, Virginia. It was designed by Robert A.M. Stern and was completed on 1 June 2009. Its owner is the College of William and Mary.

Table 2: Overview of applied zero-shot and few-shot prompts.

Issue Type	**LLaMA-7B**	**Vicuna-7B**	**GPT-3.5-Turbo**	**LoRA-7B**
	relative frequency: zero-shot / few-shot			
Inaccurate	0.60* / 0.61	0.41* / 0.48	0.13 / 0.11	0.19 / 0.17
Mistranslated	- / -	0.01* / -	- / -	- / -
Off-prompt	0.65 / -	0.27 / -	- / -	- / -
Redundant	0.23* / 0.07	0.02* / -	0.01 / 0.01	- / 0.01
Unlexicalized	0.69* / -	0.27* / -	0.07 / -	- / -

Table 3: Relative frequency of issue types for zero-shot and few-shot prompts in evaluated sample of 150 predictions with lowest averaged BLEU and METEOR scores. For values marked with "*", the relative frequency only considers generations being on-prompt.

Issue Type	Definition	Example
Inaccurate	Prediction is incomplete or factually incorrect.	`GOLD`: Binignit is a dessert that comes from the Philippines. `PRED`: Philippines is a dessert.
Mistranslated	Triples were verbalized in a language other than English.	`GOLD`: The Mexican based city of Ciudad Ayala in Morelos stands at 1147 metres above sea level and is in the PDT time zone. `PRED`: Ciudad Ayala es parte de Morelos, que es un estado en México. Ciudad Ayala tiene una elevación de 1147 metros sobre el nivel del mar [...]
Off-prompt	Prediction is unrelated to the prompt and contradicts desired output format.	`GOLD`: GMA Network is a public company. `PRED`: The input triples are: [{'object': 'Public_company', 'property': 'type', 'subject': 'GMA_Network_(company)'}].
Redundant	Unnecessary repetition of triple information.	`GOLD`: The Pontiac Rageous, a car with a coupe body style, assembled in Detroit Michigan, was first and last produced in 1997. `PRED`: 1997 Pontiac Rageous was produced in Detroit, Michigan, in 1997. It was a coupe. It was assembled in 1997. It was produced in 1997 [...] It was a coupe. It was assembled in 1997 [...]
Unlexicalized	Entities or relations are not lexicalized.	`GOLD`: The Fellowship of the Ring was followed by The Two Towers. `PRED`: The_Fellowship_of_the_Ring was followed by The_Two_Towers.

Table 4: Overview of five identified issue types with examples from generated model predictions (PRED) and human annotations (GOLD).

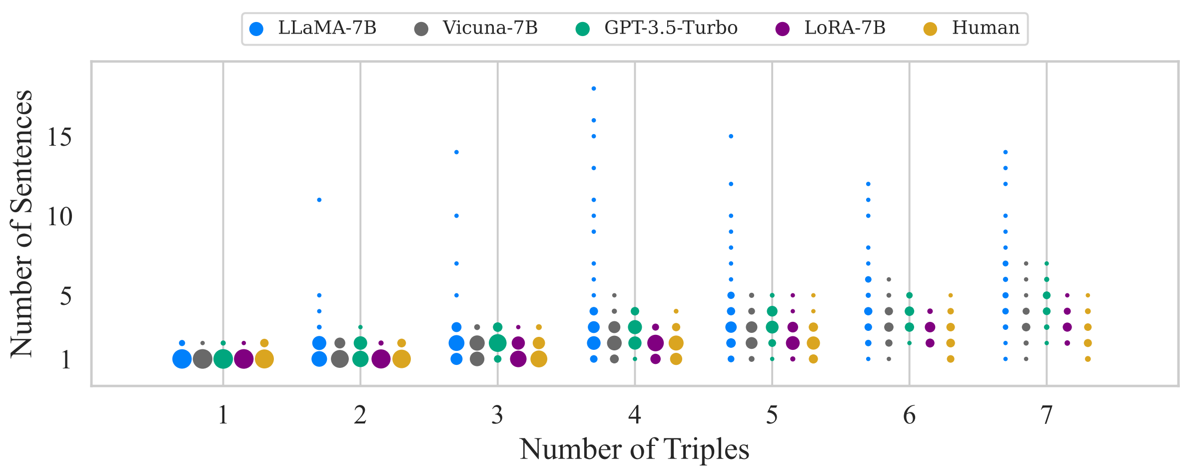

Figure 2: Distribution of model generated sentences by number of triples for few-shot models with post-processing. The size of the dots reflects the occurrence frequency. The ground-truth references are denoted as "Human".