Team SaarLST at the GEM'24 D2T Task: Symbolic retrieval substantially reduces hallucination in data-to-text generation

Mayank Jobanputra and **Vera Demberg**
{firstname}@lst.uni-saarland.de
Department of Language Science and Technology
Saarland University

Abstract

Data-to-text (D2T) generation tasks require Large Language Models (LLMs) to generate factual and faithful text from structured input. Additionally, in the counterfactual and fictional subtasks of GEM'24 shared tasks, LLMs may need to handle conflicting information from the pre-training data. Team SaarLST (Jobanputra and Demberg, 2024) introduced a few-shot retrieval-augmented generation (RAG) system centered on a symbolic retriever - `PropertyRetriever`. This work presents the analysis of the official human evaluation results from the shared task. Our system ranks first among all participating systems across all four human evaluation criteria: No-Omissions, No-Additions, Grammaticality, and Fluency. This result highlights the effectiveness of our symbolic retrieval approach in generating fluent and faithful text, even in challenging counterfactual and fictional scenarios. The human evaluation results also highlight a "reliability gap" as even state-of-the-art systems exhibit imperfections, indicating that building a reliable system for this seemingly simple task remains an open challenge.

1 System Summary

The GEM'24 shared task (Mille et al., 2024) focuses on D2T generation from RDF triplets. This shared task is primarily designed to test the faithfulness of LLMs across factual (FA), counter-factual (CFA), and fictional (FI) data. The major challenge in this task is to prevent hallucinations, where the LLM's parametric knowledge overrides the input data, and correctly inferring missing details (e.g., entity types) to generate fluent text.

Our proposed system addresses these challenges using a few-shot Retrieval-Augmented Generation (RAG) pipeline, as illustrated in our original paper (Jobanputra and Demberg, 2024). The key difference in our proposed system is a symbolic retriever - `PropertyRetriever`. Unlike dense retrievers that fetch semantically similar examples, `PropertyRetriever` creates an index of properties from the training data. At inference time, it retrieves examples that share the most properties and have a similar number of triples as the input query. This structural and property-based matching provides the generator with highly relevant stylistic and syntactic templates.

Generation Pipeline at a glance. Our pipeline consists of the following components:

- a lightweight, symbolic retriever (*e.g.*, term-similarity over RDF verbalizations) to fetch few-shot exemplars,
- in-context prompting of a general-purpose LLM for generation,
- an ensemble of two state-of-the-art open-weight LLMs: `Mixtral 8x7B` (Jiang et al., 2024) as the primary model and `Command-R` as a fallback.

Design rationale. D2T inputs (RDF triple sets) can be long-tail and compositional. We therefore prioritized an exemplar selection strategy that increases factual coverage while keeping complexity low. The literature also supports the choice of a symbolic retriever for the D2T generation task. Chang et al. (2021) showed a similar way to select relevant examples for the few-shot training. Feng et al. (2024) also used a similar retrieval mechanism for their low-resource D2T generation task.

2 Official Human Evaluation Results

Following the initial submission, the shared task organizers conducted a comprehensive human evaluation of all participating systems (Sedoc et al., 2025). The outputs are rated by human annotators on a 1-7 scale across four criteria: No-Omissions, No-Additions, Grammaticality, and Fluency.

Criterion	Team	D2T-1 (WebNLG-based) FA	CFA	FI	D2T-2 (Wikidata-based) FA	CFA	FI	Avg.
No-Omissions	SaarLST (Ours)	5.79	5.52	5.94	6.19	5.93	5.97	**5.89**
	DipInfo-UniTo	5.45	5.43	5.55	5.80	5.72	5.55	5.58
	DCU-NLG-PBN	5.49	5.25	5.57	5.46	5.41	5.38	5.43
No-Additions	SaarLST (Ours)	5.61	5.14	5.76	6.15	5.53	5.76	**5.66**
	DipInfo-UniTo	5.59	5.38	5.47	6.05	5.71	5.39	5.60
	DCU-NLG-PBN	5.56	5.10	5.48	5.48	5.08	5.16	5.31
Grammaticality	SaarLST (Ours)	6.07	5.83	5.98	6.28	6.08	6.01	**6.04**
	DipInfo-UniTo	6.01	5.68	5.81	6.12	5.95	5.55	5.85
	DCU-NLG-PBN	6.11	5.68	5.86	6.01	5.67	5.44	5.79
Fluency	SaarLST (Ours)	5.98	5.76	5.94	6.24	6.00	5.95	**5.98**
	DipInfo-UniTo	5.89	5.58	5.72	6.06	5.90	5.53	5.78
	DCU-NLG-PBN	6.04	5.60	5.81	5.92	5.63	5.46	5.74

Table 1: Official human evaluation scores (1-7 scale) for the top 3 participating systems across all subtasks. Our system (SaarLST) achieved the highest average score across every criterion.

2.1 Results and Discussion

Our system (SaarLST) ranks **first overall**, achieving the highest average score among all participating systems on every single evaluation criterion (see Table 1). This strong performance across the board validates our system's core design.

2.1.1 Faithfulness in Factual and Counterfactual Scenarios

A key goal of the shared task was to evaluate model faithfulness under challenging conditions. Our system's high scores on **No-Omissions (5.89)** and **No-Additions (5.66)** underscore the effectiveness of `PropertyRetriever` in achieving this. The retriever's ability to ground the LLM was particularly evident in the counterfactual (CFA) and fictional (FI) settings. While many systems struggle when input data conflicts with an LLM's parametric knowledge, our system maintained high faithfulness. This suggests that providing in-context examples with matching properties and structure may guide the LLM better and help prioritize the input data over its internal conflicting knowledge.

2.1.2 Grammaticality and Fluency

Beyond faithfulness, our system also excels in producing high-quality language, achieving the top scores for **Grammaticality (6.04)** and **Fluency (5.98)**. The retrieved examples provide similar discourse-level templates. This helps the LLM in structuring the information logically and connecting the individual facts into a coherent, natural-sounding paragraph. The consistency of these high scores across all six subtasks indicates that the approach is robust.

2.2 Comparative Analysis

Our system's first-place ranking becomes more insightful when viewed in the context of the other participating systems. The shared task featured a variety of approaches, with several teams employing powerful state-of-the-art LLMs, including proprietary closed-source models known for their strong generative capabilities (Mille et al., 2024).

Despite this, our system consistently outperformed all others. For example, the next-highest-performing system achieved average scores of 5.58 for No-Omissions and 5.85 for Grammaticality, compared to our 5.89 and 6.04, respectively. This outcome is particularly noteworthy given that our system was built using a symbolic retriever instead of dense retrievers in traditional RAG systems. It suggests that for faithfulness-critical tasks such as D2T, the in-context examples used to guide the model can help LLM achieve better performance, and that this effect may have a stronger influence than the capability of the base LLM itself.

The human evaluation results also indicate that simply fine-tuning an LLM on the task may yield better performance on the automated metrics, but it does not guarantee overall better performance. The tendency of LLMs to hallucinate (i.e., Addition or Omission) and fall back on parametric knowledge, especially when faced with counter-factual or fictional data, remains a noticeable limitation.

3 Discussion: a "Reliability Gap"

It is crucial to interpret these human evaluation results with appropriate skepticism for the LLM-based systems. While our system achieved the highest rank, the absolute scores ($\approx$ 6.0 out of a

possible 7) indicate that perfection is still out of reach. A score of 5.89 on 'No-Omissions,' for instance, implies that in some cases, our system did fail to convey all the provided information. This "reliability gap" suggests that even with sophisticated retrieval and generation pipelines, minor errors in faithfulness and fluency persist. These imperfections highlight the difficulty LLMs face in consistently remaining faithful to the input data. Therefore, the next challenge is not just to outperform other systems, but to fill the reliability gap.

4 Conclusion

In this work, we present the official human evaluation results for our entry in the GEM'24 D2T shared task. The results confirm that our system ranked first across all four dimensions of human judgment. A detailed comparative analysis suggests that this success stems not just from the choice of capable LLMs, but from the effectiveness of our symbolic retrieval method in ensuring better performance. This outcome provides strong evidence for the value of structured, symbolic guidance in data-to-text generation. By focusing on property-level similarity, `PropertyRetriever` provided the necessary grounding for LLMs to excel, highlighting a promising direction for future research in developing more robust and faithful NLG systems. Yet, the imperfect scores suggest a 'reliability gap' and provide an opportunity for building a truly reliable D2T generation systems.

Acknowledgments

This research is funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 389792660 – TRR 248 "Foundations of Perspicuous Software Systems"[1] and Project-ID 471607914 – GRK 2853/1 "Neuroexplicit Models of Language, Vision, and Action". We sincerely thank the GEM'24 shared task organizers for sharing the human evaluation results that helped us showcase the effectiveness of our system.

References

Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021. On training instance selection for few-shot neural 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 (Volume 2: Short Papers)*, pages 8–13, Online. Association for Computational Linguistics.

Ruitao Feng, Xudong Hong, Mayank Jobanputra, Mattes Warning, and Vera Demberg. 2024. Retrieval-augmented modular prompt tuning for low-resource data-to-text generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14053–14062, Torino, Italia. ELRA and ICCL.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mayank Jobanputra and Vera Demberg. 2024. Team-SaarLST at the GEM'24 data-to-text task: Revisiting symbolic retrieval in the LLM-age. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 92–99, Tokyo, Japan. Association for Computational Linguistics.

Simon Mille, João Sedoc, Yixin Liu, Elizabeth Clark, Agnes Axelsson, Miruna-Adriana Clinciu, Yufang Hou, Saad Mahamood, Ishmael Obonyo, and Lining Zhang. 2024. The 2024 GEM shared task on multilingual data-to-text generation and summarization: Overview and preliminary results. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

João Sedoc, Simon Mille, Miruna Adriana Clinciu, Yixin Liu, Saad Mahamood, Elizabeth Clark, Kaustubh Dhole, and Lining Zhang. 2025. The 2024 GEM shared task on multilingual data-to-text generation: English and Spanish qualitative evaluation results. In *Proceedings of the 18th International Natural Language Generation Conference: Generation Challenges*, Hanoi, Vietnam. Association for Computational Linguistics.

[1] https://perspicuous-computing.science