DCU-ADAPT-modPB at the GEM'24 Data-to-Text Task: Analysis of Human Evaluation Results

Rudali Huidrom♡†, Chinonso Cynthia Osuji♡†, Kolawole John Adebayo♡, Thiago Castro Ferreira♢, Brian Davis♡
* ADAPT Research Centre, Dublin City University, Ireland♡
Fluminense Federal University, Brazil♢
{rudali.huidrom, chinonso.osuji, brian.davis}@adaptcentre.ie

Abstract

This paper presents the official human evaluation results for DCU-ADAPT-modPB, our submission to the 2024 GEM Shared Task on Multilingual Data-to-Text Generation. The system description paper reported only automatic metrics; here we extend the analysis using the human assessments released in 2025. Annotators evaluated outputs on No-Omissions, No-Additions, Grammaticality, and Fluency across English datasets. For FA, CFA, and FI subsets, only No-Omissions scores were released, while pooled results across datasets were provided for all criteria. DCU-ADAPT-modPB achieved competitive results where it was rated above the LLM evaluation baseline and close to the human average in No-Additions, Grammaticality, and Fluency, though it lagged behind both baselines in No-Omissions. These findings demonstrate the strengths of hybrid pipelines in producing grammatical and fluent text with limited hallucination, while underscoring persistent challenges in ensuring full content coverage.

1 Introduction

Data-to-Text (D2T) generation is a long-standing goal of natural language generation (NLG), involving the production of natural language descriptions from structured data. It has applications in domains such as journalism, health reporting, business intelligence, and knowledge graph verbalisation. Despite notable progress, the field continues to face fundamental challenges: ensuring that generated text is both fluent and faithful to the input data.

The GEM benchmark (Gehrmann et al., 2021) has emerged as a standard platform for evaluating NLG systems, emphasising multilinguality, multiple tasks, and rigorous evaluation. The 2024 GEM Shared Task (Osuji et al., 2024) focused on D2T with three English datasets: factual (FA), counterfactual (CFA), and fictional (FI). These were designed to progressively increase difficulty, testing whether systems could generalise beyond in-domain factual data.

Evaluation of NLG has historically relied on automatic metrics such as BLEU (Papineni et al., 2002), ChrF++ (Popović, 2017), BERTScore (Zhang et al., 2019), and COMET (Rei et al., 2020). While efficient, these metrics are known to correlate imperfectly with human judgements, particularly for dimensions such as omissions and hallucinations (Reiter, 2018). To address this, GEM incorporates systematic human evaluation. The release of the GEM'24 human ratings in 2025 (Sedoc et al., 2025) therefore provides the most robust evidence to date of system behaviour across the new datasets.

Our system, DCU-ADAPT-modPB, adopts a hybrid pipeline design, combining symbolic structuring with LLM-based realisation. The central motivation was to mitigate hallucination by constraining the LLM to pre-structured content, while leveraging its strengths in producing fluent and grammatical sentences. This paper analyses how this design performed under human evaluation, with particular attention to the trade-off between fluency and coverage.

2 Related Work

Hybrid approaches to D2T generation have a long history, typically involving content selection, planning, and surface realisation stages (Gardent et al., 2017; Novikova et al., 2017). While such systems offer control and factual consistency, they often lag behind neural end-to-end models in terms of naturalness and fluency. Recent advances in LLMs have shifted emphasis towards end-to-end prompting or fine-tuning. These approaches achieve high

*† The first two authors made equal contributions to all aspects of the main paper. The order in which they appear in this paper was determined based on their contributions to this analysis paper.

Proceedings of the 18th International Natural Language Generation Conference: Generation Challenges, pages 44–47
October 29–November 2, 2025. ©2025 Association for Computational Linguistics

fluency but suffer from hallucinations, especially in low-resource or multilingual contexts (Maynez et al., 2020). The GEM benchmark itself has highlighted this trade-off where pipeline systems tend to avoid hallucinations but omit content, whereas end-to-end systems generate more complete but less reliable outputs (Gehrmann et al., 2021).

Within this context, our system extends the hybrid tradition. By constraining LLMs with explicit content planning, we aim to combine their strengths in form with improved factual reliability. The human evaluation results allow us to examine the extent to which this balance was achieved.

3 System Recap

The DCU-ADAPT-modPB system is a modular pipeline with three components:

- **Triple Ordering and Structuring**: Input triples were linearised and ordered with Flan-T5. This produced sentence plans that grouped semantically related triples and imposed a coherent sequence, reducing incoherence in realisation.
- **Surface Realisation**: Sentence plans were realised into natural language by prompting GPT-4 and Mistral. Prompts were designed to encourage factual faithfulness while maintaining fluency. GPT-4 contributed particularly to grammatical accuracy, while Mistral was leveraged for efficiency and diversity.
- **Translation into Target Languages**: Since English-centric LLMs currently perform best, outputs were generated first in English and then translated into Swahili and other languages using neural MT models.

This pipeline was designed to reduce hallucination while retaining fluency. We anticipated that omissions might arise during structuring, where content pruning could occur.

4 Evaluation Setup

The organisers' human evaluation (Sedoc et al., 2025) assessed system outputs on a 1–7 scale for No-Omissions, No-Additions, Grammaticality, and Fluency. For English Factual (FA), Counter Factual (CFA), and Fictional (FI) datasets, only No-Omissions scores were reported individually. For the pooled D2T-1 set (FA+CFA+FI), averages were reported for all four criteria, alongside human averages and LLM averages.

Dataset	No-Omissions
FA	5.42
CFA	5.21
FI	5.35

Table 1: No-Omissions scores for DCU-ADAPT-modPB across FA, CFA, and FI datasets.

System	No-Omis.	No-Add.	Gram.	Flu.
Human avg	**5.57**	5.73	6.33	6.25
LLM avg	5.41	5.52	6.01	5.93
DCU-ADAPT-modPB	5.33	**5.62**	**6.21**	**6.12**

Table 2: English D2T-1 pooled results (FA+CFA+FI). Human ratings averaged across all criteria.

5 Results

5.1 Per-dataset No-Omissions

The following results are per-dataset no-omissions results (see Table 1):

- On FA, DCU-ADAPT-modPB scored 5.42.
- On CFA, the score dropped to 5.21, reflecting the increased difficulty of counterfactual reasoning.
- On FI, the system achieved 5.35, consistent with its conservative bias under more creative inputs.

These results suggest that DCU-ADAPT-modPB is effective at minimising unsupported additions to the input, but it frequently under-generates by omitting relevant content. The tendency towards omission is especially pronounced in the CFA setting, where altered input facts increase the difficulty of maintaining full coverage.

5.2 Pooled Results (D2T-1)

Across FA, CFA, and FI combined, DCU-ADAPT-modPB performed strongly in No-Additions, Grammaticality, and Fluency, outperforming the LLM baseline and approaching the human average. In No-Omissions, however, it lagged behind both baselines. See Table 2.

6 Discussion

The results highlight a clear profile. DCU-ADAPT-modPB excels in producing grammatical and fluent text with few hallucinations, as reflected in its superior scores on No-Additions, Grammaticality, and Fluency. However, its conservative design results in lower No-Omissions, especially in CFA, where the system struggled to cover perturbed inputs.

When compared with baselines, DCU-ADAPT-modPB performs close to human averages in linguistic quality, but below both humans and LLMs in coverage. This illustrates the persistent coverage–accuracy trade-off: systems that constrain generation to reduce hallucination often omit input content, whereas more expansive systems cover more but risk errors.

Upon manual inspection of the intermediate outputs produced during the content ordering and structuring stages, it was observed that the `Flan-T5` model occasionally omitted some input triples even before the surface realisation stage. Although no quantitative calculation of omission rate has yet been conducted, these preliminary observations suggest that older encoder–decoder models such as `Flan-T5` are more prone to partial content loss when handling complex or lengthy input sets. In contrast, newer and larger models (e.g., GPT-4, Claude, or Mistral-7B) appear to exhibit fewer such omissions during generation, likely due to their improved contextual reasoning and long-context consistency.

These findings also raise broader methodological issues. The fact that human-authored references do not dominate all criteria suggests that annotation guidelines reward certain forms of fidelity and conciseness differently from natural human variation.

These findings also raise broader methodological issues. The fact that human-authored references do not dominate all criteria suggests that annotation guidelines reward certain forms of fidelity and conciseness differently from natural human variation. This reinforces calls for multi-dimensional evaluation frameworks that account for pragmatic adequacy, diversity, and user needs in addition to surface fidelity.

Future work should address omissions directly, for example through reinforcement learning from human feedback (Christiano et al., 2017) or direct preference optimisation (Rafailov et al., 2023), which could encourage models to balance coverage with linguistic quality.

7 Conclusion

We presented the human evaluation results for DCU-ADAPT-modPB, our submission to GEM'24. The system outperformed the LLM baseline and closely matched human averages in No-Additions, Grammaticality, and Fluency, but underperformed in No-Omissions. This reflects the strengths and weaknesses of hybrid pipelines: they deliver reliable, readable text with minimal hallucination, yet often sacrifice completeness. Addressing omissions remains the critical challenge for future D2T research.

Limitations

Our analysis is limited to the English datasets, as human evaluation was not released for other languages.

Ethics Statement

This work carries minimal risk. It reports analysis of human evaluation results under controlled conditions.

References

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *10th International Conference on Natural Language Generation*, pages 124–133. ACL Anthology.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, and 1 others. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254*.

Chinonso Cynthia Osuji, Rudali Huidrom, Kolawole John Adebayo, Thiago Castro Ferreira, and Brian Davis. 2024. Dcu-adapt-modpb at the gem'24 data-to-text generation task: Model hybridisation for pipeline data-to-text natural language generation. In *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 66–75.

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.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.