DCU-ADAPT-modPB at the GEM'24 Data-to-Text Generation Task: Model Hybridisation for Pipeline Data-to-Text Natural Language Generation

Chinonso Cynthia Osuji[♡†], **Rudali Huidrom**[♡†], **Kolawole John Adebayo**[♡],
Thiago Castro Ferreira[◇], **Brian Davis**[♡]
ADAPT Research Centre, Dublin City University, Ireland[♡]
aiXplain, USA[◇]
chinonso.osuji@adaptcentre.ie rudali.huidrom@adaptcentre.ie
kolawole.adebayo@adaptcentre.ie thiago@aixplain.com
brian.davis@adaptcentre.ie

Abstract

In this paper, we present our approach to the GEM Shared Task at the INLG'24 Generation Challenges, which focuses on generating data-to-text in multiple languages, including low-resource languages, from WebNLG triples. We employ a combination of end-to-end and pipeline neural architectures for English text generation. To extend our methodology to Hindi, Korean, Arabic, and Swahili, we leverage a neural machine translation model. Our results demonstrate that our approach achieves competitive performance in the given task.

1 Introduction

The GEM 2024 Shared Task (Mille et al., 2024) aims to advance summarisation and data-to-text (D2T) generation, with a particular focus on enhancing multilingual capabilities. The D2T task (Reiter and Dale, 1997) involves generating coherent natural language text from structured data in the form of Wikidata and WebNLG datasets, which are organised as triples consisting of a subject, predicate, and object. The goal of the tasks is to comprehensively evaluate and improve the ability of systems to interpret and generate text from RDF triples, assess their general knowledge, and produce texts in factual (FA), counterfactual (CFA), and fictional (FI) scenarios.

The dominance of English in D2T generation presents a considerable challenge, highlighting the need for research to support effective multilingual generation, particularly for languages with diverse morphological structures and distinct word order characteristics. The GEM 2024 Shared Task addresses this challenge by including English alongside other languages such as Chinese, German, Russian, Spanish, Korean, Hindi, Swahili, and Arabic, which are low-resource in the D2T setting. This task aims to enhance the adaptability and robustness of different systems across varied linguistic frameworks for text generation from structured data.

In this submission, we focus on the D2T generation aspect of the task using the WebNLG dataset (Castro Ferreira et al., 2020). Our approach combines end-to-end and pipeline neural architectures to generate English text, while also fine-tuning a state-of-the-art open-source Flan-T5 and Mistral-7B large language models (LLMs) for generating text in low-resource languages. Our approach aims to further the understanding of how various architectures can be optimised for multilingual D2T generation. Our methodology demonstrates competitive performance and contributes substantial insights and advancements to the field of multilingual D2T generation. The code and results are available[1].

2 Related Work

The field of data-to-text generation has undergone significant transformations, evolving from traditional pre-neural approaches that relied on handcrafted rules, templates, and statistical models (Reiter and Dale, 1997; Erdem et al., 2022) to modern deep learning architectures. These advanced models are trained to identify and replicate the relationships between structured data and its corresponding textual outputs. The introduction of end-to-end systems, particularly pre-trained language models (PLMs), has substantially improved the processing of textual sequences in data-to-text tasks (Kale and Rastogi, 2020; Ribeiro et al., 2021). However, despite their advanced capabilities, these systems often struggle with content selection and maintaining fidelity due to the opaqueness and complexity inherent to deep learning models and the data-to-text generation task(Moryossef et al., 2019).

[†] The first two authors made equal contributions to all aspects of the work, the order in which they appear was decided arbitrarily.

[1] https://github.com/NonsoCynthia/GEM2024_ST

Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, pages 66–75
September 23 –27, 2024.

A recent example of methodological advancement in this field is showcased in the 2023 WebNLG Shared Task on Low Resource Languages, where many participants employed NLG+MT (Natural Language Generation plus Machine Translation) pipeline approach (Cripwell et al., 2023). For instance, some participants implemented systems which generate English text from RDF graphs using a PLM fine-tuned on the WebNLG 2020 dataset, followed by translation into various languages using a machine translation (MT) model (Aditya Hari et al., 2023; Kumar et al., 2023). This approach showcases the potential of combining NLG and MT models for effective multilingual data-to-text generation.

Similarly, Lorandi and Belz (2023) proposed a novel approach that utilises large language models (GPT-3.5 and GPT-4) for prompt-based generation without additional training. They tested two methods: direct generation in under-resourced languages and generation in English followed by translation using Google Translate. In our research, we build upon these methodologies and incorporate a 3-stage pipeline neural architecture, as in Figure 1, inspired by Ferreira et al. (2019). However, we modify the approach by integrating only the first two stages of ordering and structuring, followed by the final stage of surface realisation. This approach aims to optimise the use of large language models for multilingual data-to-text generation.

3 Methodology

In this section, we outline the methodologies employed to address the generation challenge for the languages English (en), Hindi (hi), Korean (ko), Arabic (ar), and Swahili (sw). Our experimental setup is as follows:

3.1 Data

We utilised the enhanced WebNLG dataset (Castro Ferreira et al., 2018) for fine-tuning the ordering and structuring stages in the intermediate phases of the pipeline neural architecture. For fine-tuning the Mistral7b model, we used the WebNLG'17 dataset (Gardent et al., 2017). Finally, we evaluate the performance of the fine-tuned models using the GEM 2024 Shared Task D2T dataset, which encompasses factual, fictional, and counterfactual domains, each containing 1779 RDF triple sets.

3.2 System Description

The GEM 2024 Shared Task focuses on summarisation and data-to-text (D2T) generation, with a particular emphasis on multilingual capabilities. For this task, only testing data is provided, consisting of three parallel datasets: Factual (FA), Counterfactual (CFA), and Fictional (FI). The FA dataset uses original triples from WebNLG'20 data (Castro Ferreira et al., 2020) and Wikidata (Vrandečić and Krötzsch, 2014), while the CFA dataset replaces entities in the factual dataset with similar-class entities, e.g., by swapping person names, dates, etc. The FI dataset substitutes entities in the factual dataset with fabricated entities generated by large language models (LLMs). Our work concentrates exclusively on data-to-text generation of triples from WebNLG.

Pipeline Neural Architecture: We designed a pipeline neural architecture, depicted in Figure 1, which leverages the fine-tuned Flan-T5-*large* model (Chung et al., 2022) to perform ordering and structuring tasks on the enhanced WebNLG 2017 dataset (Castro Ferreira et al., 2018). The Flan-T5 model is initially fine-tuned separately for ordering and structuring tasks using a subset of the enhanced WebNLG dataset. As shown in Figure 2, the pipeline architecture takes test set triples (FA, CFA, FI) as input and passes them through the ordering model to determine their verbalisation sequence. The ordered triples are then mapped to their corresponding entities (subjects and objects values) and fed into the structuring model. The structuring model organises the entities into coherent sentences, marking sentence boundaries with [SNT] and [/SNT] tags, while ensuring accurate entity mappings. Predicates serve as pointers during this process, linking to their respective triples after generation.

Finally, for surface realisation, we integrated prompt-based models, including Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) and GPT-4 Turbo (Ye et al., 2023; Achiam et al., 2023). The structured outputs are fed into these prompt-based models to generate the final text. The overall workflow is presented in Figure 2.

Parameter Efficient Instruction Fine-Tuning: Our second setup employs parameter efficient fine-tuning (PEFT) (Houlsby et al., 2019) for instruction tuning of the selected models. Specifically, we utilise LORA (Hu et al., 2021), which inte-

	BLEU ↑	METEOR ↑	ChrF++ ↑	TER ↓	BERT_P ↑	BERT_R ↑	BERT_F1 ↑
StructGPT4	**49.80**	0.40	0.655	**0.450**	**0.958**	0.953	**0.955**
GPT4	42.823	**0.418**	**0.677**	0.548	0.948	**0.957**	0.952
Mistral	37.552	0.378	0.623	0.559	0.943	0.949	0.945
StructMistral	35.493	0.353	0.584	0.578	0.940	0.941	0.940
FinetunedMistral	31.070	0.29	0.513	0.630	0.913	0.916	0.914

Table 1: Automatic metrics results of our systems for factual (FA) English test set. Bold and underlined results denote the best and the second best ones respectively.

	FACTUAL				
	Arabic	**Hindi**	**Korean**	**Swahili**	**English**
StructGPT4	0.499	0.425	0.581	0.612	0.629
GPT4	0.546	0.478	**0.633**	**0.627**	**0.636**
Mistral	**0.558**	0.445	0.608	0.613	0.625
StructMistral	0.498	**0.615**	0.581	0.612	0.615
FinetunedMistral	0.498	0.276	0.433	0.574	0.551
	COUNTERFACTUAL				
	Arabic	**Hindi**	**Korean**	**Swahili**	**English**
StructGPT4	0.511	0.406	0.576	0.567	0.49
GPT4	**0.551**	**0.448**	**0.613**	0.571	**0.518**
Mistral	0.519	0.415	0.584	0.580	0.471
StructMistral	0.479	0.374	0.542	**0.581**	0.441
FinetunedMistral	0.308	0.239	0.372	0.556	0.254
	FICTIONAL				
	Arabic	**Hindi**	**Korean**	**Swahili**	**English**
StructGPT4	0.508	0.408	0.589	0.554	**0.499**
GPT4	0.137	0.062	0.180	**0.564**	0.108
Mistral	**0.530**	**0.428**	**0.602**	0.559	0.484
StructMistral	0.494	0.397	0.575	0.563	0.460
FinetunedMistral	0.300	0.231	0.369	0.532	0.238

Table 2: COMET metrics results of our systems for FA, CFA and FI test set for all the languages. Bold and underlined results denote the best and the second best ones respectively.

grates trainable adapters in the form of low-rank decomposition matrices into chosen layers of a transformer model. To enhance the diversity of our training data, we designed a template that produces 10 rewritten instructions for each original instruction. These re-written instructions are worded differently, but convey the same meaning or action trigger, allowing the fine-tuned model to align more robustly to varied instructions and improve its ability to generalise to new, unseen inputs. We use the the WebNLG'17 corpus (Gardent et al., 2017) for the model fine-tuning. We then combine the fine-tuned model with the base model, leveraging both the specialised fine-tuning and the broad knowledge inherent from pretraining. This composite model is tested with 5 examples from the WebNLG corpus, along with our newly created dataset.

In-Context Learning: In our final setup, we utilised the in-context learning (Zhao et al., 2023; Yang et al., 2024) capabilities of the selected models, namely Mistral7b, and GPT-4, for text generation tasks. We performed few-shot prompting using

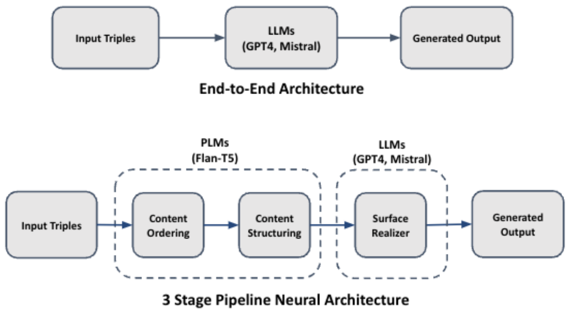

Figure 1: System Description.

five triples randomly selected from the WebNLG corpus. The prompt designs used in our experiments are presented in Appendix A.

3.3 Machine Translation Model

The English outputs generated by the systems described in Section 3.2 were translated into Hindi, Korean, Arabic, and Swahili using specialised machine translation models. For the translation of Korean, Arabic, and Swahili, we utilised the open-source Command-R-Plus model developed by Cohere (Üstün et al., 2024). Specifically, we utilised the 4-bit quantised version which is available on the HuggingFace model hub[2]. The translation into Hindi was performed using the IndicTrans2 model (Gala et al., 2023), which is also an open-source transformer-based multilingual NMT model specifically trained for all 22 officially recognised Indic languages. Our selection of the two multilingual models was based on their open-source availability and their relative performance in the languages covered in our experiments. We conducted preliminary limited testing to evaluate their performance by having native language speakers assess the quality of the translated text. Their feedback informed our decision to use these translation models for our experiments.

[2]https://huggingface.co/CohereForAI/c4ai-command-r-plus-4bit

4 Results

In our results' naming convention, "Struct" denotes the pipeline architecture system that utilises structured triples for generation. "FinetunedMistral" refers to the fine-tuned Mistral-7B-Instruct system, while systems without these acronyms represent direct generation using the base models within the end-to-end architecture.

The results from the evaluation in Table 1 provide valuable insights into the strengths and weaknesses of the different models across various automatic metrics within the English language in the FA dataset. StructGPT4 achieved the highest scores in BLEU (49.80), TER (0.45), BERT_P (0.958), and BERT_F1 (0.955) for English. Following this, GPT4 consistently emerges as the most versatile and high-performing model, excelling in a wide range of languages (Arabic, Hindi, Korean, Swahili, and English) and domains (FA, CFA, FI). For instance, in the FA English test set, GPT4 achieves top scores in METEOR (0.418), ChrF++ (0.677), and BERT_F1 (0.952), underscoring its ability to produce translations that are both semantically accurate and closely aligned with reference texts.

Furthermore, we employed the COMET metric (Rei et al., 2020), a neural evaluation model specifically designed to predict quality scores for translations. COMET is known for demonstrating a strong correlation with human judgement and is capable of performing reference-less evaluations. This capability makes COMET particularly well-suited for assessing our results in non-English languages within the FA dataset, as well as for all languages in the CFA and FI datasets, where reference translations are not yet available. The results of our evaluation using COMET are presented in Table 2. The results indicate that GPT-4 consistently performs well, particularly in the FA and CFA datasets, achieving the highest scores in English (0.636 for FA, 0.518 for CFA) and in several other languages (see Table 2). However, GPT-4 struggles in the FI dataset, especially in Arabic, Hindi, and Korean, with scores as low as 0.137 in Arabic. Mistral shows strong performance across all datasets, particularly excelling in the FI dataset, where it achieves the highest scores in Arabic (0.530), Hindi (0.428), and Korean (0.602). StructGPT4 also performs well, leading in the FI dataset with a score of 0.499 in English, and shows strong results in other datasets, especially in Arabic and Korean. StructMistral is competitive in Swahili, particularly in the CFA dataset (0.583), but generally ranks second in most other cases. In contrast, FinetunedMistral underperforms across all languages and datasets, with notably low scores, such as 0.254 in English for the CFA dataset. Overall, GPT-4 and Mistral emerge as the top-performing models for the COMET metrics, but their effectiveness varies depending on the dataset and language, highlighting the importance of context in model performance.

5 Analysis and Discussion

In this analysis, we highlight the factors which may have contributed to the varying performances of the models in our experiments.

First, the underlying architecture and training data play a critical role. We observe that our GPT4-based systems benefits from extensive training on a large and diverse dataset, which likely contributes to its consistent performance across different languages and domains. The robustness of its architecture allows it to handle a wide range of tasks effectively. However, we observed a decline in performance within the FI dataset. Upon manual inspection, we found that the system generated text with the correct entities but often rejected certain entity claims in the dataset, leading to its overall poor performance in this category.

Second, the fine-tuning process and the nature of the tasks significantly influence performance. StructGPT4, for instance, is fine-tuned with a focus on specific tasks (i.e., ordering and structuring) requiring precision and the handling of complex or nuanced content, which explains its superior performance in BLEU and TER, especially in FA English text generation.

Third, language-specific optimisations or model adaptations can lead to better performance in certain languages. Mistral shows strong results in Korean and Swahili, which may indicate that it has been trained or optimised for these specific languages, allowing it to outperform GPT4 and StructGPT4 in these contexts.

Fourth, the evaluation metrics themselves might favour certain models depending on how they align with the strengths of each model. For example, StructGPT4 performs better in BLEU and TER, metrics that emphasise precision and reduced errors, while GPT4 excels in METEOR and ChrF++, which also account for semantic accuracy and fluency.

These factors highlight the importance of selecting models based on the specific requirements of the task, considering not only the general capabilities of the model but also how well it has been optimised or fine-tuned for particular languages and tasks. To fully harness the aggregate benefits of the various factors influencing the performance of models as identified in our experiment, future work should focus on conducting a comprehensive exploration of each aspect. This may involve:

- **Experimental Design Optimisation:** Investigating different architectural designs, such as combining structured and prompt-based approaches, to identify the most effective methods for enhancing model performance.

- **Fine-tuning Strategies:** Exploring fine-tuning techniques that can better balance the retention of learned general capabilities and adaptation to specific tasks, thereby minimising the risk of overfitting and improving model generalisation.

- **Dataset Selection:** Examining the impact of training data on model performance by comparing the performance of these models when finetuned with canonical datasets from multiple GEM and WebNLG competitions, thereby gaining insights on dataset diversity and size on model adaptation and generalisation for D2T generation tasks.

- **Evaluation Methods:** Enhancing evaluation methodologies by integrating both automatic and human evaluations, ensuring a more accurate and nuanced assessment of model performance. This may involve developing new metrics that can better capture the subtleties of generated text in the context of D2T tasks.

6 Conclusion and Future Directions

In conclusion, this paper presents the methodologies and automatic evaluation results of our submission to the GEM 2024 tasks. The evaluation results highlight the strengths of different models across various metrics and languages. StructGPT4 stands out in producing precise translations with fewer errors, especially in English, outperforming GPT4 in metrics like BLEU and TER. GPT4, however, proves to be the most versatile and high-performing model across multiple languages and domains, excelling in METEOR, ChrF++, BERT_F1, and COMET metrics, although it shows limitations in generating text within the FI task.

Mistral demonstrates strong performance in languages such as Korean, Hindi, and Arabic, particularly within the FI task, while StructMistral excels in Swahili CFA tasks. These findings suggest that while GPT4 is the most reliable general-purpose model, StructGPT4, due to its incorporation of task splitting and pipelining, is better suited for tasks requiring minimal errors, high accuracy, and attention to detail. Meanwhile, Mistral and StructMistral offer valuable performance in specific applications, indicating their potential for specialised use cases.

In order to gain a more comprehensive understanding of our systems' performance, we look forward to the availability of human evaluation results, which will provide valuable insights and enable us to draw further conclusions. Moreover, we plan to further explore the impact of advanced fine-tuning methods with preference-based learning, such as recent state-of-the-art frameworks like DPO (Rafailov et al., 2024), KTO (Ethayarajh et al., 2024), SPPO (Wu et al., 2024) and the REINFORCE (Ahmadian et al., 2024) preference optimisation. These methods have shown promise in improving model alignment and generation performance, and we believe they could be valuable additions to our existing systems.

We will also investigate the possible impact of data selection and prompt engineering methods on optimising our existing systems. Studies, for example in (Shen, 2024; Liu et al., 2024) have shown that carefully selecting and preparing high-quality data for LLM finetuning often leads to improvement in model performance. This is because high-quality data allows the model to learn from relevant and accurate examples, which is crucial for fine-tuning the model's parameters and achieving optimal performance.

Lastly, we are keen on investigating the development of an end-to-end framework that encompasses ordering, structuring, and text generation collectively. This would allow us to streamline our pipeline and potentially improve the overall performance of our systems.

Ethics Statement

We adhered to the structure of the ARR responsible research checklist. The risk associated with this study was minimal.

Acknowledgments

Osuji's work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223; and ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University under Grant No 13/RC/2106_P2. Huidrom's work on this project was supported by the Faculty of Engineering and Computing, DCU, via a PhD grant. Adebayo's contribution on this work was supported by Enterprise Ireland's CareerFit-Plus Co-fund and the European Union's Horizon 2020 research and innovation programme Marie Skłodowska-Curie Grant No. 847402. We would like to thank Prof. Anya Belz for her kind guidance and discussions throughout the course of this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kancharla Aditya Hari, Bhavyajeet Singh, Anubhav Sharma, and Vasudeva Varma. 2023. WebNLG challenge 2023: Domain adaptive machine translation for low-resource multilingual RDF-to-text generation (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 93–94, Prague, Czech Republic. Association for Computational Linguistics.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

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, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Liam Cripwell, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, and William Soto Martinez. 2023. The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 55–66, Prague, Czech Republic. Association for Computational Linguistics.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

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. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 552–562.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea

Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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, Dublin, Ireland. Association for Computational Linguistics.

Nalin Kumar, Saad Obaid Ul Islam, and Ondrej Dusek. 2023. Better translation + split and generate for multilingual RDF-to-text (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 73–79, Prague, Czech Republic. Association for Computational Linguistics.

Ziche Liu, Rui Ke, Feng Jiang, and Haizhou Li. 2024. Take the essence and discard the dross: A rethinking on data selection for fine-tuning large language models. *arXiv preprint arXiv:2406.14115*.

Michela Lorandi and Anya Belz. 2023. Data-to-text generation for severely under-resourced languages with GPT-3.5: A bit of help needed from Google Translate (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 80–86, Prague, Czech Republic. 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.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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 and Robert Dale. 1997. Building applied natural language generation systems.

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.

Ming Shen. 2024. Rethinking data selection for supervised fine-tuning. *arXiv preprint arXiv:2402.06094*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Prompt Design

Table 4 illustrates our prompt design for English with five examples each for with and without structured data. We report two types of prompts, one

Input Triples:
[TRIPLE] Bananaman broadcastedBy BBC [/TRIPLE] [TRIPLE] Bananaman creator John_Geering [/TRIPLE] [TRIPLE] Bananaman firstAired "1983-10-03" [/TRIPLE] [TRIPLE] Bananaman lastAired "1986-04-15" [/TRIPLE] [TRIPLE] Bananaman starring Tim_Brooke-Taylor [/TRIPLE]

Ordering Output:
broadcastedBy firstAired lastAired creator starring

Input Triples after mapping:
[TRIPLE] Bananaman broadcastedBy BBC [/TRIPLE] [TRIPLE] Bananaman firstAired "1983-10-03" [/TRIPLE] [TRIPLE] Bananaman lastAired "1986-04-15" [/TRIPLE] [TRIPLE] Bananaman creator John_Geering [/TRIPLE] [TRIPLE] Bananaman starring Tim_Brooke-Taylor [/TRIPLE]

Structuring Output:
[SNT] broadcastedBy firstAired lastAired [/SNT] [SNT] creator starring [/SNT]

Input Triples after mapping and removing the [TRIPLE] start and [TRIPLE] end tags:
[SNT] Bananaman broadcastedBy BBC, Bananaman firstAired "1983-10-03", Bananaman lastAired "1986-04-15" [/SNT] [SNT] Bananaman creator John_Geering, Bananaman starring Tim_Brooke-Taylor [/SNT]

Surface Realizer Output:
Bananaman was shown on the BBC, first airing on October 3, 1983 and the final broadcast being April 15, 1986. It was created by John Geering and starred Tim Brooke Taylor.

Figure 2: Pipeline Neural Architecture Outputs

Factual Dataset Result							
	Models	**0**	**1**	**2**	**3**	**4**	**Average**
English	**StructGPT4**	0.8132	0.8096	0.7654	0.3805	0.3781	0.629
	GPT4	0.8189	0.815	0.7713	0.3874	0.3851	0.636
	Mistral	0.8035	0.8005	0.7583	0.383	0.3808	0.625
	StructMistral	0.7855	0.7838	0.7425	0.3832	0.3809	0.615
	FinetunedMistral	0.6909	0.6884	0.6525	0.3619	0.3596	0.551
Arabic	**StructGPT4**	0.6228	0.6208	0.5919	0.3317	0.3296	0.499
	GPT4	0.684	0.6821	0.6509	0.357	0.3552	0.546
	Mistral	0.6817	0.6807	0.65	0.3902	0.3884	0.558
	StructMistral	0.6046	0.6043	0.5755	0.3521	0.3496	0.497
	FinetunedMistral	0.605	0.6048	0.5758	0.3521	0.3497	0.498
Hindi	**StructGPT4**	0.5061	0.5083	0.4859	0.3122	0.3102	0.425
	GPT4	0.5847	0.5854	0.5588	0.3307	0.3291	0.478
	Mistral	0.5395	0.542	0.5177	0.3145	0.313	0.445
	StructMistral	0.4818	1.4841	0.4649	0.3232	0.3211	0.615
	FinetunedMistral	0.3196	0.3209	0.2101	0.2665	0.2646	0.276
Korean	**StructGPT4**	0.6828	0.6817	0.6549	0.4426	0.4409	0.581
	GPT4	0.7473	0.7466	0.7196	0.4777	0.4759	0.633
	Mistral	0.7205	0.7196	0.6925	0.4555	0.4541	0.608
	StructMistral	0.6704	0.6705	0.6466	0.4602	0.4581	0.581
	FinetunedMistral	0.4701	0.4696	0.4572	0.385	0.3832	0.433
Swahili	**StructGPT4**	0.6513	0.6504	0.6389	0.5602	0.5593	0.612
	GPT4	0.6671	0.6663	0.6544	0.5742	0.5733	0.627
	Mistral	0.652	0.6514	0.6402	0.5621	0.5614	0.613
	StructMistral	0.6485	0.6482	0.6379	0.5639	0.5629	0.612
	FinetunedMistral	0.6033	0.6026	0.5935	0.5365	0.5356	0.574

Table 3: Factual dataset COMET results of the individual reference texts (0, 1, 2, 3, & 4) for evaluation.

for GPT4 model and the other for the Mistral-7B-Instruct model.

Table 5 presents our prompt design for translating English to Arabic, Korean and Swahili using command-r-plus-4bit model from Cohere AI. We provide five examples each for the respective languages.

System instruction	"You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being socially unbiased and safe. If you're unsure about an answer, it's okay to skip it, and please ensure not to provide incorrect information. Additionally, responses should be concise and informative."
User instruction	"I would like you to generate a fluent and concise summaries or text in English based on the triples provided. Below you may find examples of the input triples and the expected summary outputs. Do not omit any triple information in the text or include any information that cannot be directly inferred from the given triples."
Data examples	**1:** *'Input'*: 'Uruguay leader Tabaré_Vázquez, Uruguay leader Raúl_Fernando_Sendic_Rodríguez, Alfredo_Zitarrosa deathPlace Montevideo, Montevideo country Uruguay', *'Output'*: 'Alfredo Zitarrosa died in Montevideo, Uruguay which is led by Raúl Fernando Sendic Rodríguez and Tabaré Vázquez.', **2:** *'Input'*: 'Angola_International_Airport location Ícolo_e_Bengo, Ícolo_e_Bengo country Angola, Angola_International_Airport cityServed Luanda, Ícolo_e_Bengo isPartOf Luanda_Province, Angola_International_Airport elevationAboveTheSeaLevelInMetres 159', *'Output'*: 'Angola International Airport is located at Ícolo e Bengo in Luanda province, Angola. The Airport is situated 159 meters above sea level and serves the city of Luanda.', **3:** *'Input'*: 'United_Petrotrin_F.C. ground Palo_Seco, Akeem_Adams club Trinidad_and_Tobago_national_under-20_football_team, Akeem_Adams club United_Petrotrin_F.C.', *'Output'*: 'Akeem Adams, who plays for the Trinidad and Tobago national under-20 football team previously played for United Petrotrin FC whose ground is at Palo Seco.', **4:** *'Input'*: 'William_Anders selectedByNasa 1963, William_Anders nationality United_States, William_Anders birthDate "1933-10-17", William_Anders occupation Fighter_pilot, William_Anders birthPlace British_Hong_Kong, William_Anders mission Apollo_8', *'Output'*: 'The United States fighter pilot William Anders was born in British Hong Kong on the 17th of October, 1933. In 1963, he was chosen by NASA and became a crew member on Apollo 8.', **5:** *'Input'*: "Dead_Man's_Plack location England, England ethnicGroup British_Arabs, England capital London, Dead_Man's_Plack dedicatedTo Æthelwald,_Ealdorman_of_East_Anglia, England language Cornish_language, England religion Church_of_England, Dead_Man's_Plack material Rock_(geology)", *'Output'*: "The capital of England is London where we can find the Dead Man's Plack which is made of stone. The Plack is dedicated to Æthelwald, Ealdorman of East Anglia. Cornish language is spoken in England and it has an established religion called the Church of England. One of the ethnic groups found in that country is the British Arabs."
Source	Input triple(s) from the test set. E.g. Andra_(singer) genre Rhythm_and_blues, Andra_(singer) background "solo_singer", Rhythm_and_blues derivative Disco
(Structured) Data examples	**1:** *'Input'*: "[SNT] [TRIPLE] Atatürk_Monument_(İzmir) material 'Bronze' [/TRIPLE] [TRIPLE] Atatürk_Monument_(İzmir) inaugurationDate '1932-07-27' [/TRIPLE] [/SNT] [SNT] [TRIPLE] Atatürk_Monument_(İzmir) location Turkey [/TRIPLE] [TRIPLE] Turkey capital Ankara [/TRIPLE] [TRIPLE] Turkey largestCity Istanbul [/TRIPLE] [/SNT] [SNT] [TRIPLE] Turkey leaderName Ahmet_Davutoğlu [/TRIPLE] [TRIPLE] Turkey currency Turkish_lira [/TRIPLE] [/SNT]", *'Output'*: "The Atatürk Monument is a bronze monument inaugurated on 27th July, 1932, in Izmir. It is found in Turkey, a country which has Ankara as its capital and Istanbul as its largest city. The leader of Turkey is called Ahmet Davutoğlu, and the currency is the Turkish lira.", **2:** *'Input'*: "[SNT] [TRIPLE] Turkey capital Ankara [/TRIPLE] [TRIPLE] Turkey largestCity Istanbul [/TRIPLE] [/SNT] [SNT] [TRIPLE] Turkey leader Ahmet_Davutoğlu [/TRIPLE] [TRIPLE] Turkey currency Turkish_lira [/TRIPLE] [/SNT] [SNT] [TRIPLE] Atatürk_Monument_(İzmir) location Turkey [/TRIPLE] [/SNT]", *'Output'*: "The capital of Turkey is Ankara, although the largest city is Istanbul. The leader of Turkey is Ahmet Davutoglu and the currency is known as the Turkish lira. The Ataturk monument is located within the country.", **3:** *'Input'*: "[SNT] [TRIPLE] Antwerp_International_Airport cityServed Antwerp [/TRIPLE] [TRIPLE] Antwerp country Belgium [/TRIPLE] [TRIPLE] Belgium leaderName Philippe_of_Belgium [/TRIPLE] [TRIPLE] Belgium language French_language [/TRIPLE] [/SNT]", *'Output'*: "Antwerp is served by Antwerp International Airport and is a popular tourism destination in Belgium where the leader is Philippe of Belgium and the French language is spoken.", **4:** *'Input'*: "[SNT] [TRIPLE] AWH_Engineering_College state Kerala [/TRIPLE] [TRIPLE] AWH_Engineering_College country India [/TRIPLE] [TRIPLE] AWH_Engineering_College established 2001 [/TRIPLE] [/SNT] [SNT] [TRIPLE] India river Ganges, India largestCity Mumbai [/TRIPLE] [/SNT] [SNT] Kerala leaderName Kochi [/TRIPLE] [/SNT]", *'Output'*: "The AWH Engineering College in Kerala, India was established in 2001. The Ganges is a river in India and its largest city is Mumbai. The leader of Kerala is Kochi.", **5:** *'Input'*: "[SNT] [TRIPLE] Atlanta country United_States [/TRIPLE] [TRIPLE] United_States capital Washington_D.C. [/TRIPLE] [/SNT] [SNT] [TRIPLE] United_States ethnicGroup Asian_Americans [/TRIPLE] [/SNT]", *'Output'*: "Atlanta is in the United States whose capital is Washington, D.C. Asian Americans are an ethnic group in the U.S."
Source	Input triple(s) from the test set. E.g. [SNT] Bananaman broadcastedBy BBC, Bananaman firstAired "1983-10-03", Bananaman lastAired "1986-04-15" [/SNT] [SNT] Bananaman creator John_Geering, Bananaman starring Tim_Brooke-Taylor [/SNT]
Our Prompt(s)	**GPT-4**: {User instruction}\n Examples:{Data examples}\n Input: {Source}\n Output:\n **Mistral7b**: <s>[INST] «SYS» {System instruction}\n{User instruction}\nExamples:{Data examples}«/SYS»\nInput: {source}\nOutput:\n[/INST]

Table 4: Prompt design for English. The first data examples are for direct prompt-based experiments and the latter are for model hybridisation experiments.

Target language	Arabic, Korean, Swahili
Instruction	"Translate the following English language text to {tgt_lang} language text. Provide only the translation. Follow the example below. ######"
Data Examples	**1:** '*Input*': 'Alfredo Zitarrosa died in Montevideo, Uruguay which is led by Raúl Fernando Sendic Rodríguez and Tabaré Vázquez.', '*Arabic*': .توفي ألفريدو زيتاروزا في مونتيفيديو، أوروغواي التي يقودها راؤول فرناندو سينديتش رودريغيز وتاباري فاسكيز، '*Korean*': "영국의 수도는 런던으로, 돌로 만든 데드맨스 플랙(Dead Man's Plack)을 찾을 수 있습니다. Plack은 East Anglia의 Ealdorman인 Æthelwald에게 헌정되었습니다. 영국에서는 콘월어가 사용되며 영국 교회라는 종교가 확립되어 있습니다. 그 나라에서 발견되는 인종 그룹 중 하나는 영국계 아랍인입니다.", '*Swahili*': "Mji mkuu wa Uingereza ni London ambapo tunaweza kupata Plack ya Dead Man ambayo imetengenezwa kwa mawe. Plack imejitolea kwa Æthelwald, Ealdorman wa East Anglia. Lugha ya Cornish inazungumzwa nchini Uingereza na ina dini iliyoanzishwa inayoitwa Kanisa la Anglikana. Moja ya makabila yanayopatikana katika nchi hiyo ni Waarabu wa Uingereza.", **2:** '*Input*': 'Angola International Airport is located at Ícolo e Bengo in Luanda province, Angola. The Airport is situated 159 meters above sea level and serves the city of Luanda.', '*Arabic*': .يقع مطار أنغولا الدولي في ايكولو ايبينغو في مقاطعة لواندا، أنغولا. يقع المطار على ارتفاع ٩٥١ متراً فوق مستوى سطح البحر ويخدم مدينة لواندا، '*Korean*': "앙골라 국제공항은 앙골라 루안다 지방의 이콜로 에 벤고에 위치해 있습니다. 공항은 해발 159미터에 위치해 있으며 루안다 시에 서비스를 제공합니다.", '*Swahili*': "Uwanja wa ndege wa Kimataifa wa Angola uko Ícolo e Bengo katika jimbo la Luanda, Angola. Uwanja wa ndege upo mita 159 juu ya usawa wa bahari na unahudumia jiji la Luanda.", **3:** '*Input*': 'Akeem Adams, who plays for the Trinidad and Tobago national under-20 football team previously played for United Petrotrin FC whose ground is at Palo Seco.', '*Arabic*': .أكيم آدامز، الذي يلعب لصالح منتخب ترينيداد وتوباغو لكرة القدم تحت ٠٢ سنة، سبق له اللعب مع نادي يونايتد بيتروترين لكرة القدم الذي يقع ملعبه في بالو سيكو، '*Korean*': "트리니다드토바고 20세 이하 축구 국가대표팀에서 뛰고 있는 아킴 아담스는 팔로세코를 연고지로 하는 유나이티드 페트로트린 FC에서 선수 생활을 했습니다.", '*Swahili*': "Akeem Adams, anayechezea timu ya taifa ya vijana ya Trinidad na Tobago ya soka ya vijana chini ya umri wa miaka 20 hapo awali aliichezea United Petrotrin FC ambayo uwanja wake ni Palo Seco.", **4:** '*Input*': 'The United States fighter pilot William Anders was born in British Hong Kong on the 17th of October, 1933. In 1963, he was chosen by NASA and became a crew member on Apollo 8.', '*Arabic*': ".٨ ولد الطيار المقاتل الأمريكي ويليام أندرس في هونغ كونغ البريطانية في ٧١ أكتوبر ٣٣٩١. وفي عام ٣٦٩١، تم اختياره من قبل وكالة ناسا وأصبح أحد أفراد طاقم أبولو ، '*Korean*': "미국 전투기 조종사 윌리엄 앤더스는 1933년 10월 17일 영국령 홍콩에서 태어났어요. 1963년 NASA에 발탁되어 아폴로 8호의 승무원이 되었습니다.", '*Swahili*': "Rubani wa kivita wa Marekani William Anders alizaliwa Uingereza Hong Kong tarehe 17 Oktoba, 1933. Mnamo 1963, alichaguliwa na NASA na kuwa mwanachama wa wafanyakazi kwenye Apollo 8.", **5:** '*Input*': "The capital of England is London where we can find the Dead Man's Plack which is made of stone. The Plack is dedicated to Æthelwald, Ealdorman of East Anglia. Cornish language is spoken in England and it has an established religion called the Church of England. One of the ethnic groups found in that country is the British Arabs.", '*Arabic*': .عاصمة إنجلترا هي لندن حيث يمكننا العثور على نصب ديدمان بلاك تذكاري المصنوع من الحجر. المقام مخصص للملك إيثلووف، زعيم و قائد من شرق أنجليا. يتم التحدث باللغة الكورنية في إنجلترا ولها دين راسخ يسمى كنيسة إنجلترا. إحدى المجموعات العرقية الموجودة في ذلك البلد هي العرب البريطانيين، '*Korean*': "영국의 수도 런던에는 돌로 만든 데드맨의 플랙이 있습니다. 이 플랙은 이스트 앵글리아의 에델발드에게 헌정되어 있어요. 영국에서는 콘월어를 사용하며 영국 국교회라는 종교가 확립되어 있습니다. 이 나라에서 발견되는 인종 그룹 중 하나는 영국 아랍인입니다.", '*Swahili*': "Mji mkuu wa Uingereza ni London ambapo tunaweza kupata Plack ya Dead Man ambayo imetengenezwa kwa mawe. Plack imejitolea kwa Æthelwald, Ealdorman wa East Anglia. Lugha ya Cornish inazungumzwa nchini Uingereza na ina dini iliyoanzishwa inayoitwa Kanisa la Anglikana. Moja ya makabila yanayopatikana katika nchi hiyo ni Waarabu wa Uingereza."
Source	System outputs from GPT-4 or Mistral7b. E.g. Aaron Turner, a post-metal singer, started his active years in 1995. He is associated with the band Twilight.
Our Prompt	{instruction} $\backslash n$Examples:{examples}$\backslash n$Input: {source}$\backslash n$Output:$\backslash n$

Table 5: Prompt design for translation of English to Arabic, Korean and Swahili using the command-r-plus-4bit model from Cohere AI.