Scaling Up Data-to-Text Generation to Longer Sequences: A New Dataset and Benchmark Results for Generation from Large Triple Sets

Chinonso Cynthia Osuji[♡], **Simon Mille**[♡], **Ornait O'Connell**[♡], **Thiago Castro Ferreira**[◇], **Anya Belz**[♡], **Brian Davis**[♡]
[♡] ADAPT Research Centre, Dublin City University, Ireland
[◇] Fluminense Federal University, Brazil
firstname.lastname@adaptcentre.ie, thiago.ferreira@gmail.com

Abstract

The ability of LLMs to write coherent, faithful long texts from structured data inputs remains relatively uncharted, in part because nearly all public data-to-text datasets contain only short input-output pairs. To address these gaps, we benchmark six LLMs, a rule-based system and human-written texts on a new long-input dataset in English and Irish via LLM-based evaluation. We find substantial differences between models and languages.

1 Introduction

Large and Very Large Language Models (LLMs/VLLMs) have set a new standard for natural language generation (NLG), producing fluent and accurate text across a host of applications such as text simplification (Dou et al., 2023), image captioning (Mokady et al., 2021), and question answering (Erdem et al., 2022; Akermi et al., 2020). Yet the very capabilities that allow for excellent short-form output can fail to scale to the extended output sequences demanded by tasks such as summarization, creative writing, and detailed question answering (Bai et al., 2024a,b; An et al., 2024; Kuratov et al., 2024).

This challenge has given rise to an emerging focus on long-form generation, which requires models to preserve topical focus, logical progression, and factual consistency across lengthy passages or entire documents (Liu et al., 2024, 2025). Consequently, the field has seen the development of specialized 'long-output' LLMs that are specifically tuned to handle these longer generative tasks (Wu et al., 2025). As inputs or required outputs extend beyond the model's trained context window, generation quality deteriorates markedly (Yu et al., 2025). Even within the context window, performance declines for longer input lengths (Liu et al., 2023a) and extended outputs (Yang et al., 2025), leading to reduced coherence, greater factual inaccuracies, and an increased incidence of hallucinations (Li et al., 2024). These issues underscore a critical gap in the ability of current models to maintain logical consistency and relevance over extended output sequences in data-to-text tasks.

Despite the significant advancements and availability of 'long-context' LLMs, the Data-to-Text (D2T) generation field continues to rely heavily on outdated datasets and benchmarks featuring predominantly short input-output pairs. In addition, such datasets can lead to data leakage, where LLMs have been trained on data closely resembling or identical to the test samples, undermining the validity of evaluations (Zhou et al., 2023; Li et al., 2024; Balloccu et al., 2024). This reliance presents a critical limitation, as these datasets are increasingly insufficient for rigorously challenging contemporary LLMs, given their substantial capacity for managing extensive contexts.

In this paper, our contributions are the following: (i) method and code for collecting long-input data for D2T generation, along with a dataset of 537 long inputs; (ii) English and Irish system outputs for these 537 inputs, as well as human-written texts for a subset of 21 inputs; and (iii) a preliminary evaluation of output quality in both languages using an LLM-as-judge approach and diversity metrics. All code, data and results are available at https://github.com/mille-s/Build_KGs_entities.

2 Related Work

While the broader field of NLP has seen recent work on long-context LLMs, primarily focused on improving model capabilities for reasoning and understanding extensive input contexts (Jiang et al., 2023; Qwen et al., 2025; Wu et al., 2025), research specifically on long-form output generation within D2T remains nascent. In the field of D2T generation, benchmark datasets have been pivotal in driving significant progress by providing structured data for model training and evaluation. However, the rapidly evolving capabilities of contemporary

Proceedings of the 18th International Natural Language Generation Conference, pages 810–822
October 29–November 2, 2025.

(Very) Large Language Models, or (V)LLMs, such as OpenAI's GPT models (Achiam et al., 2023; OpenAI, 2024), Meta's Llama models (Touvron et al., 2023b; Dubey et al., 2024), Google's Gemini (Team et al., 2024), and Anthropic's Claude (Anthropic, 2024), highlight a growing need for datasets featuring significantly longer token sequences. Existing foundational D2T datasets, while valuable, consistently exhibit relatively short average output lengths:[1] E2E (22.67 tokens) (Novikova et al., 2017), WebNLG (22.69 tokens) (Gardent et al., 2017; Castro Ferreira et al., 2020), ToTTo (17.4 tokens) (Parikh et al., 2020), and WeatherGOV (28.70 tokens) (Liang et al., 2009), WikiBio (26.1 tokens) (Lebret et al., 2016). While some datasets like RotoWire (337.10 tokens) (Wiseman et al., 2017) offer longer outputs, the latter typically necessitate additional content selection components which complicates its application to direct D2T generation. The short input-output token lengths of the majority of D2T datasets severely limit their utility for effectively evaluating the long-context handling and extended-text generation capabilities of modern LLMs.

Another significant limitation is the dominance of English datasets, which restricts the applicability of D2T models to multilingual and low-resource scenarios. Recent initiatives aimed at broadening linguistic inclusivity emphasize the importance of developing benchmarks that explicitly include low-resource languages (Cripwell et al., 2023; Mille et al., 2024). Our research directly addresses this by incorporating a low-resource language, Irish, into the evaluation framework of our proposed dataset.

3 Experimental Setup

We created a new dataset of long-context data-to-text inputs (Section 3.1), generated outputs with a range of different (V)LLMs (Section 3.2), and evaluated the outputs along with human-written texts via LLM-as-judge assessment (Section 3.3).

3.1 Dataset

We compiled a small dataset of DBpedia input triple sets of varying sizes by (i) collecting the 602 entities of the `City` and `Person` categories from the GREC shared task (Belz et al., 2009); (ii) for each entity, querying DBpedia for all triples in which the entity appears either as Subject or Object, limiting the search to all properties used in the WebNLG shared tasks, which resulted in the collection of 258,067 triples for the 602 entities; (iii) curating the triple sets as follows: (a) filtering out properties identified as incorrect, including second to nth occurrence of a property that can only have one value (e.g. *birthDate*) and properties frequently misused on DBpedia, (b) allowing for a maximum of 3 instances of the same property with the same Subject or Object,[2] and (c) selecting only the triple sets with at least 8 triples (i.e., triples sets of larger size than in WebNLG). The dataset comprises 537 triple sets with sizes ranging from 8 to 69 triples, for an average of 24.6 triples per input; a sample data point is provided in Appendix A. We collected reference texts for a subset of inputs (see Section 3.3).

3.2 Systems

We evaluated a suite of recent LLMs of different sizes for data-to-text generation: GPT-4.1 (OpenAI et al., 2024) and Claude-sonnet-3.7 (Anthropic, 2024), which both support English and Irish, QWEN3-32B (Team, 2025), DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI et al., 2025) and LLaMA2-13B (Touvron et al., 2023a) for English, and UCCIX (Tran et al., 2024), a domain-adapted, finetuned version of LLaMA2-13B for Irish, based on the assumption that domain-specific adaptation could yield measurable improvements. Finally, we also used the off-the-shelf rule-based system FORGe (Mille et al., 2023), developed for both languages on the WebNLG data. Details about model use are provided in Appendix B.

3.3 Evaluation

Scoring of outputs. LLMs have been shown to be very reliable evaluators for various NLG tasks such as dialogue generation (Liu et al., 2023b), machine translation (Kocmi and Federmann, 2023), and in particular data-to-text generation (Huidrom and Belz, 2025). For assessing the quality of the different outputs, we ran and averaged the scores assigned by two different LLMs, GPT-o3 and Claude-sonnet-3.7, two of the best models currently available. We used the same four dimensions as in WebNLG, text quality is assessed in its own right (*Grammaticality* and *Fluency*), and with respect to the input (*No-omissions* and *No-additions*), on

[1] Some of the statistics were extracted from the original papers, and the remainder come from Lin et al. (2023).

[2] For instance, it is often the case that a city is the Object of several hundreds of location properties, which would result in very unnatural texts with endless coordinations.

a scale from 1 (lowest quality) to 7 (highest quality); definitions are provided in the prompt in Appendix C, and detailed results in Appendix D.[3]

Data sampling and human references. In order to assess the capability of LLMs to assess the quality of long outputs in a D2T context, we also evaluate human-written texts. A native English and Irish speaker (an author) wrote texts starting from the input triple sets, following a set of simple instructions and a list of definitions of the different properties used in the data; these outputs are labeled Hum-1.0 in Figure 1. Because it is a time-consuming task, we randomly sampled twenty-one data points evenly across seven different triple size ranges: 8–9, 10–19, 20–29, 30–39, 40–49, 50–59, and 60–69 (three data points per size range). The resulting 21 triple sets have an average size of 35.4 triples (compared to a maximum of 7 in WebNLG).

Truncated texts for sanity check. We also added three truncated versions of the human texts in order to control whether models are able to detect missing contents, with respectively 10% (Hum-0.9), 30% (Hum-0.7) and 50% (Hum-0.5) of the sentences removed from the end of original texts.

4 Results and Analysis

4.1 Results of LLM-as-judge evaluation

Figure 1 shows the evaluation results; each plot shows the scores of each output for one language-dimension pair and for the different input sizes. The main observations are the following.

Very Large Language Models (GPT, Claude, Qwen) seem to be able to generate good texts from long-input structured data. The degradation of the scores on longer inputs across all criteria is quite moderate, and is less visible than on smaller models (LLaMA, DeepSeek, UCCIX), rule-generated and human-written outputs. However, it is more challenging for the VLLMs (in particular GPT) to not add content than to not omit content: the VLLM *No-additions* scores are lower than the *No-omissions* scores, and *No-additions* is the only dimension for which Human and FORGe have similar or better scores than VLLMs in both languages.

VLLM scores are overall slightly lower in Irish than in English, while for human-written and rule-generated texts, it is the opposite, with higher scores for Irish than for English. While the *No-omissions* evaluation looks similar in both languages, in Irish, there is less difference between VLLM outputs and the other outputs (except UCCIX) for *Grammaticality* and *Fluency*, and for No-Additions, GPT outputs score below all other outputs (but UCCIX) across all sizes. This probably reflects the greater difficulty of generating text in an under-resourced language like Irish.

VLLMs seem to be able to pass the sanity checks as evaluators. They are able to detect semantic content mismatches in both languages. In terms of *No-omissions*, the more a human output is truncated, the lower the score (Hum-1.0 > Hum-0.9 > Hum-0.7 > Hum-0.5). We noticed that as the inputs get larger, FORGe can be unable to generate some triples in Irish, and this is also reflected in the Irish *No-omissions* plot with a clearly descending curve. Additionally, rule-based systems such as FORGe traditionally score high on semantic accuracy metrics (*No-additions* and *No-additions*) and low on Fluency, as is the case in Figure 1.

However, **it is likely that the VLLMs used as judge are positively biased towards their own outputs**, a phenomenon already documented (Panickssery et al., 2024). In our evaluation plots, VLLMs consistently assign higher *Fluency* and *Grammaticality* scores to their own outputs than to human-written texts, particularly in English. This pattern is unlikely to correspond to genuine differences in their quality and instead suggests that VLLMs may exhibit a self-serving bias when acting as evaluators. A human assessment of the outputs (and of the LLM judgments) in both languages would be needed to draw more solid conclusions.

4.2 Diversity analysis of the outputs

In order to get a deeper understanding of the texts produced by the different systems, we used several metrics proposed in (Shaib et al., 2025) to assess output diversity.[4] For our analysis, we consider only the output texts used in the evaluation (i.e. 21 texts per system, see Section 4.1). We use the following **four** metrics as well as the average text length in words: (i) **The N-Gram Diversity score** considers n-grams of size 1 to 4 (**NGD-1to4**). This score is calculated as the ratio of unique n-grams to all n-grams in the 21 texts. A higher number indicates a higher overall lexical diversity in the

[3]We also ran LlaMA2 and DeepSeek-R1-Distill-Llama and report the numbers in App. D; they were discarded because the models assigned multiple maximum overall semantic accuracy scores, which is very unlikely to happen on our dataset.

[4]Please refer to the paper for exact formulas, and to our GitHub repository for the code used.

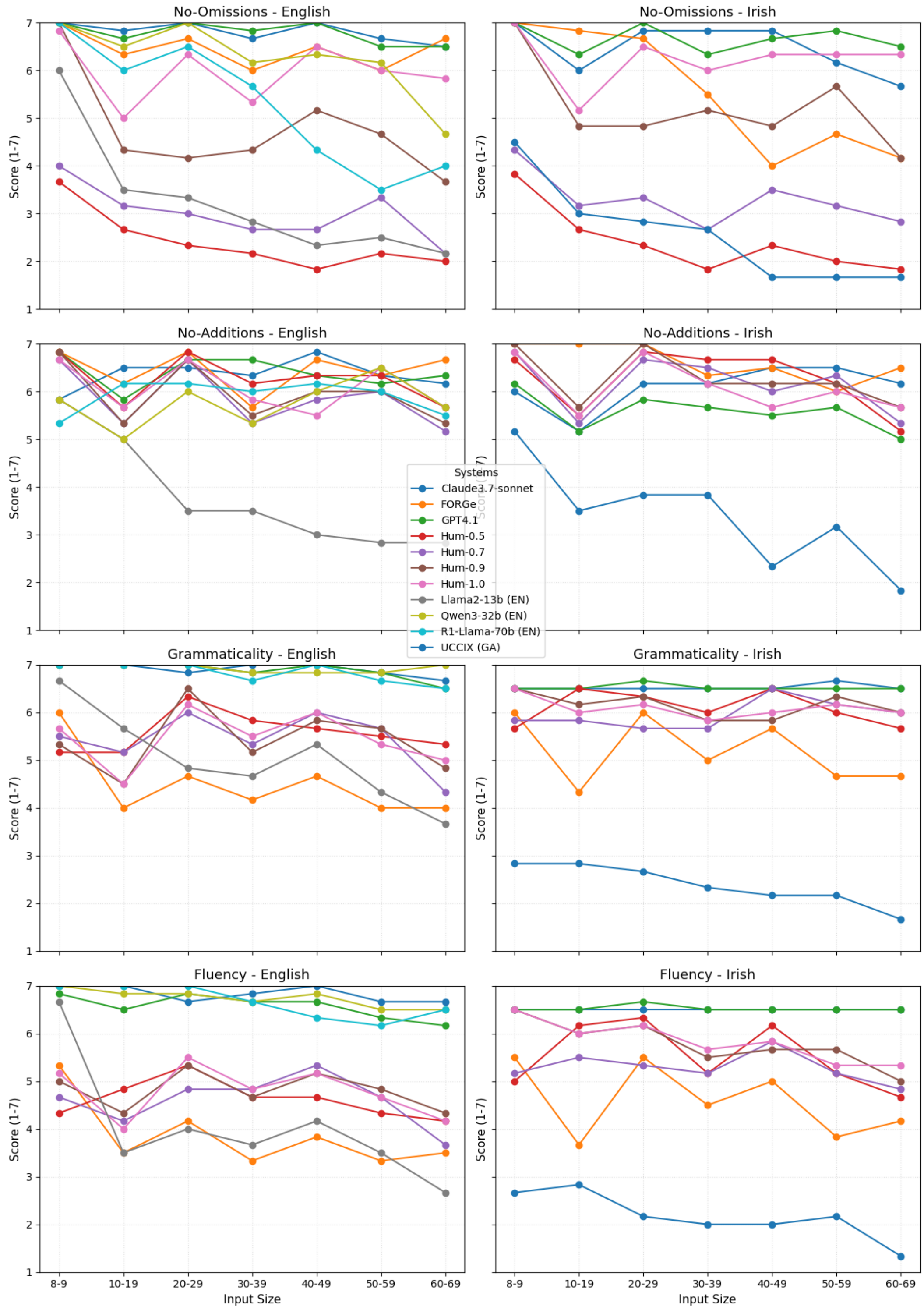

Figure 1: System performance across criteria and input sizes; left: English (EN), right: Irish (GA).

English

System	AvgLenW	NGD-1to4 (↑)	CompRate (↓)	SelfRep-4 (↓)	ChamDist (↑)
Hum-1.0	191.3	2.872	2.603	3.117	0.484
Claude3.7-sonnet	315.1	2.854	2.836	3.678	0.504
FORGe	**184.9**	**2.869**	**2.594**	3.129	0.498
GPT4.1	279.3	2.698	2.831	3.737	0.503
Llama2-13b	305.5	1.588	4.365	2.278	0.**486**
Qwen3-32b	261.7	2.921	2.692	2.939	0.488
R1-Llama-70b	235.3	2.867	2.687	**3.111**	0.479

Irish

System	AvgLenW	NGD-1to4 (↑)	CompRate (↓)	SelfRep-4 (↓)	ChamDist (↑)
Hum-1.0	189.5	3.030	2.555	2.969	0.451
Claude3.7-sonnet	263.1	**2.946**	2.801	**2.973**	0.423
FORGe	**169.7**	2.843	**2.638**	3.486	0.494
GPT4.1	320.1	2.765	2.906	3.683	**0.432**
UCCIX	263.1	1.849	4.349	1.351	0.515

Table 1: Diversity analysis in English and Irish: Average length in words (AvgLenW), N-Gram Diversity score (NGD-1to4), Compression Rate (CompRate), Self-Repetition score (SelfRep-4) and Chamfer Distance (ChamDist). The direction of the arrows indicates whether high (up) or low (down) scores mean more diversity. In **bold**, the score(s) closest to the score of the human-written text (first row); underlined, the best absolute score in a column.

outputs. (ii) **The Compression Rate** (**CompRate**) is the ratio between the size of the 21 texts compressed with gZip and the size before compression. A lower compression rate is an indicator of a higher overall (sub)string diversity. (iii) **The Self-Repetition score** considers 4-grams only (**SelfRep-4**), calculated using the number of other texts in which 4-grams of one text are repeated and averaging for all 21 texts. A lower SelfRep-4 score indicates lower n-gram repetition rate across outputs, hence higher inter-sentence diversity. (iv) **The Chamfer Distance** (**ChamDist**) is computed as the mean pairwise cosine distance between texts, using Qwen3-Embed-0.6B[5] for embeddings, to quantify semantic diversity. Higher scores indicate greater meaning variation between system outputs, while lower scores suggest greater meaning homogeneity.

Overall, FORGe has the most similar diversity scores to human-written texts in English, while in Irish, Claude3.7-Sonnet is most similar to human texts. LLama2-13B (EN) and UCCIX (GA) score much lower than others according to NGD-1to4 and CompRate, while they do good in absolute terms (but quite different from human texts) according to SelfRep-4, which indicates that their texts are more (and possibly overly) different from each other but individually less diverse. All other systems are generally close to human-written texts for these metrics. In terms of semantic diversity, since the texts are about different entities but with overlapping properties, it is expected that ChamDist, whose values ranges from 0 to 2, is neither high nor too low. Finally, only FORGe produces texts close to the length of human texts, while (V)LLMs output texts usually over 50% longer word-wise; since text length can affect the diversity scores (e.g. longer texts give more opportunities for n-gram overlap), these should be interpreted with caution.

[5] https://huggingface.co/Qwen/Qwen3-Embedding-0.6B

5 Conclusions

We developed a method for creating data for long-input data-to-text generation, compiled a new DBpedia-based dataset, ran several systems on the inputs and assessed text quality using an LLM-as-judge approach. We conclude that some (V)LLMs seem able to generate long texts from structured data and to evaluate their quality along several dimensions, although there may be biases in the evaluation of LLMs by LLMs. In Irish, Claude appears to be the best model, while a fine-tuned version of UCCIX does not perform well; it is unclear whether this is due to limitations of the fine-tuning data (the WebNLG'23 Irish data is automatically translated from English) or of the pretrained model itself, as indicated by LlaMA-2's low English scores. In future work, we will add filters to optimise the selection of high-quality triples for dataset creation. We will also carry out human assessment of the text quality and of the collected ratings so as to gain further insights into the use of LLMs for long-context D2T and its evaluation.

6 Limitations

Due to the cost and difficulty of performing an human evaluation on such a dataset, the texts are only assessed using an LLM-as-judge approach. Although the evaluation results are in line with what one can expect using several sanity checks, the reliability of the evaluation needs to be confirmed. In addition, creating human-written texts for such long inputs is time consuming, only 21 texts per system were assessed. Additional texts would be desirable for drawing more solid conclusions (for each of the 21 inputs of 35.4 triples on average, it took over 1 hour to write a reference text in both English and Irish).

The input data was collected automatically, and despite carefully checking its quality and adding mechanisms to limit the presence of wrong triples, a small proportion of triples with bad object values are still found in the data, e.g. *Ibn al-Tilmidh – occupation – Baghdad* (there should be a job title instead of "Baghdad"), or *Al-Mustansir II – predecessor – Baghdad* (here a person name was expected as the Object of *predecessor*). Finally, although we release system outputs for possible further analyses and reproduction studies, these outputs are not meant to be used for learning.

Ethics Statement

We use LLM-based methods in our experiments, and at present, it is uncertain what data has been used to train them, especially proprietary models such as GPT and Claude. The texts they produced and the assessments they provided may reflect biases, potentially posing a risk of harm to users.

Acknowledgments

Osuji's contribution was supported by Research Ireland Centre for Research Training in Artificial Intelligence (CRT-AI) under Grant No. 18/CRT/6223.

Mille's contribution was funded by the European Union under the Marie Skłodowska-Curie grant agreement No 101062572 (M-FleNS), and by the Irish Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media via the eSTÓR project. Our work has also benefited more generally from being carried out within the research environment of the ADAPT SFI Centre, funded by Science Foundation Ireland through the SFI Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106_P2.

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*.

Imen Akermi, Johannes Heinecke, and Frédéric Herledan. 2020. Transformer based natural language generation for question-answering. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 349–359.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2024. Claude. Accessed: 2025-06-06.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2024a. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Anja Belz, Eric Kow, Jette Viethen, and Albert Gatt. 2009. The GREC main subject reference generation challenge 2009: Overview and evaluation results. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*, pages 79–87, Suntec, Singapore. Association for Computational Linguistics.

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.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Yao Dou, Philippe Laban, Claire Gardent, and Wei Xu. 2023. Automatic and human-ai interactive text generation.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Erkut Erdem, Barbara Plank, Albert Gatt, Emiel Krahmer, Mandar Sharma, Ajay Gogineni, Naren Ramakrishnan, Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. *INLG 2017 - 10th International Natural Language Generation Conference, Proceedings of the Conference*, 298:124–133.

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.

Rudali Huidrom and Anja Belz. 2025. Ask me like i'm human: Llm-based evaluation with for-human instructions correlates better with human evaluations than human judges. In *Proceedings of the 4th Table Representation Learning Workshop*, pages 98–108.

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.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *24th Annual Conference of the European Association for Machine Translation*, page 193.

Yury Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 1203–1213.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can long-context language models understand long contexts? In *Proceedings* of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Percy Liang, Michael Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore. Association for Computational Linguistics.

Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2023. A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, Yuanxing Zhang, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, Zhejian Zhou, Ruijie Zhu, Junlan Feng, Yang Gao, Shizhu He, Zhoujun Li, Tianyu Liu, Fanyu Meng, Wenbo Su, Yingshui Tan, Zili Wang, Jian Yang, Wei Ye, Bo Zheng, Wangchunshu Zhou, Wenhao Huang, Sujian Li, and Zhaoxiang Zhang. 2025. A comprehensive survey on long context language modeling. *CoRR*, abs/2503.17407.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024. Longgenbench: Long-context generation benchmark. *arXiv preprint arXiv:2410.04199*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Simon Mille, Elaine Uí Dhonnchadha, Lauren Cassidy, Brian Davis, Stamatia Dasiopoulou, and Anja Belz. 2023. Generating irish text with a flexible plug-and-play architecture. In *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 25–42.

Simon Mille, João Sedoc, Yixin Liu, Elizabeth Clark, Agnes Johanna Axelsson, Miruna Adriana Clinciu, Yufang Hou, Saad Mahamood, Ishmael Nyunya 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 Natural Language Generation Conference: Generation Challenges*, pages 17–38, Tokyo, Japan. Association for Computational Linguistics.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

OpenAI. 2024. Hello gpt-4o | openai. Accessed: 2025-06-06.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David

Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1173–1186.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2025. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team. 2025. Qwen3 technical report.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D Nguyen. 2024. Uccix: Irish-excellence large language model. In *ECAI 2024*, pages 4503–4506. IOS Press.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Yuhao Wu, Yushi Bai, Zhiqing Hu, Shangqing Tu, Ming Shan Hee, Juanzi Li, and Roy Ka-Wei Lee. 2025. Shifting long-context llms research from input to output. *arXiv preprint arXiv:2503.04723*.

Joonho Yang, Seunghyun Yoon, Hwan Chang, Byeongjeong Kim, and Hwanhee Lee. 2025. Hallucinate at the last in long response generation: A case study on long document summarization. *arXiv preprint arXiv:2505.15291*.

Yijiong Yu, Yongfeng Huang, Zhixiao Qi, and Zhe Zhou. 2025. Training with “paraphrasing the original text” teaches llm to better retrieve in long-context tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25751–25759.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don’t make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

A Sample data point

Figure 2 below shows a triple set collected with the method described in Section 3.1.

B Details about model use for generating texts.

We fine-tuned Llama2-13B and the UCCIX model using the WebNLG dataset (Cripwell et al., 2023) in both English and Irish. All experiments employed parameter-efficient LoRA adaptation (Hu et al., 2021) for data-to-text generation. Training was conducted on the ADAPT HPC cluster equipped with NVIDIA A100 GPUs (80GB), allowing for efficient large-scale fine-tuning.

All models were accessed via their respective APIs: OpenAI (GPT-4.1), Anthropic (claude-3-7-sonnet-latest), and Hugging Face (Llama 2 13B, UCCIX). For each evaluation, the temperature was set to 1; the prompts are shown in Table 2 below.

C Template prompt for LLM-as-judge evaluation

For all our LLM-based evaluations, we used the following prompt, only changing the “Triple Set” and “Text” values at the end according to the evaluated data point (shown here with a short input for more clarity; we invoked the models directly via their official APIs using the code in `https://github.com/mille-s/GEM24_EvalLLM`):

In this task, you will evaluate the quality of the Text in relation to the given Triple Set. How well does the Text represent the Triple Set? You will be given four specific Dimensions to evaluate against:

Dimensions:""" No-Omissions: ALL the information in the Triple Set is present in the Text. No-Additions: ONLY information from the Triple Set is present in the Text. Grammaticality: The Text is free of grammatical and spelling errors. Fluency: The Text flows well and is easy to read; its parts are connected in a natural way."""

Important note on No-Omissions and No-Additions: some Triple Set/Text pairs contain non-factual information and even fictional names for people, places, dates, etc. Whether there are omissions and/or additions in a Text is NOT related to factual truth, but instead is strictly related to the contents of the input Triple Set. Important note on Grammaticality and Fluency: for Grammaticality and Fluency you do not need to consider the input Triple Set; only the intrinsic quality of the Text needs to be assessed.

You need to provide the scores ranging from 1 (indicating the lowest score) to 7 (indicating the highest score) for each of the dimensions and a short justification for each score in the following JSON format: "No-Omissions": "Justification": "", "Score": "", "No-Additions": "Justification": "", "Score": "", "Grammaticality": "Justification": "", "Score": "", "Fluency": "Justification": "", "Score": "" .

Make sure to read thoroughly the Triple Set and the English Text below, and assess the four Dimensions using the instructions and template above.

Triple Set: """Marcus_Aurelius HasChild Fadilla; Marcus_Aurelius StudentOf Alexander_of_Cotiaeum; Marcus_Aurelius Spouse Faustina_the_Younger; Marcus_Aurelius PositionHeld Roman_emperor; Marcus_Aurelius PlaceOfDeath Vindobona""" Text: Marcus Aurelius has Fadilla as child, he supervised Alexander of Cotiaeum and is married to Faustina the Younger. He plays in Roman emperor and passed away in Vindobona.

D Detailed LLM-as-judge evaluation results

Figures 3-10 show the details of the LLM-as-judge evaluation. Note that despite not supporting Irish, LLaMA and DeepSeek models are able to detect omissions in Irish, and particularly low levels of quality of the text in its own right (*Fluency* and *Grammaticality* of UCCIX and FORGe) and of *No-additions* (UCCIX). Regarding Claude-3.7 and GPT-o3, Claude generally gives higher ratings than GPT for the semantic accuracy criteria (*No-omissions*, *No-additions*), while for the other two criteria, it depends on the language: GPT tends to give higher scores in English, while Claude tends to give higher scores in Irish. These differences will be the subject of further analysis in future work.

```xml
<?xml version="1.0" ?>
<benchmark>
  <entries>
    <entry category="City" eid="161" shape="(X (X) (X) (X) (X))" shape-type="sibling" size="39">
      <originaltripleset>
        <otriple>Quetta | areaMetro | 3501000000.0</otriple>
        <otriple>Quetta | areaTotal | 3501.0</otriple>
        <otriple>Quetta | areaCode | 081</otriple>
        <otriple>Quetta | elevation | 1679.448</otriple>
        <otriple>Quetta | populationTotal | 1001205</otriple>
        <otriple>Quetta | postalCode | 87xxx</otriple>
        <otriple>Quetta | utcOffset | +05:00</otriple>
        <otriple>Quetta | country | Pakistan</otriple>
        <otriple>Quetta | timeZone | Pakistan_Standard_Time</otriple>
        <otriple>Quetta | type | Metropolis</otriple>
        <otriple>Provincial_Assembly_of_Balochistan | location | Quetta</otriple>
        <otriple>Provincial_Disaster_Management_Authority_(Balochistan) | location | Quetta</otriple>
        <otriple>Quetta_Development_Authority | location | Quetta</otriple>
        <otriple>Qamar_Zaman | birthPlace | Quetta</otriple>
        <otriple>Qasim_Suri | birthPlace | Quetta</otriple>
        <otriple>Qazi_Faez_Isa | birthPlace | Quetta</otriple>
        <otriple>Hussain_Ali_Yousafi | deathPlace | Quetta</otriple>
        <otriple>Meena_Keshwar_Kamal | deathPlace | Quetta</otriple>
        <otriple>Safdar_Kiyani | deathPlace | Quetta</otriple>
        <otriple>Mohammad_Anwar_Khan_Durrani__Tenure__1 | state | Quetta</otriple>
        <otriple>Baluchistan_(Chief_Commissioner's_Province) | capital | Quetta</otriple>
        <otriple>Baluchistan_Agency | capital | Quetta</otriple>
        <otriple>Quetta_Gladiators | city | Quetta</otriple>
        <otriple>Quetta_International_Airport | city | Quetta</otriple>
        <otriple>Sardar_Bahadur_Khan_Women's_University | city | Quetta</otriple>
        <otriple>10th_Princess_Mary's_Own_Gurkha_Rifles | garrison | Quetta</otriple>
        <otriple>4th_(Quetta)_Division | garrison | Quetta</otriple>
        <otriple>XII_Corps_(Pakistan) | garrison | Quetta</otriple>
        <otriple>Balochistan_cricket_team | ground | Quetta</otriple>
        <otriple>Provincial_Disaster_Management_Authority_(Balochistan) | headquarter | Quetta</otriple>
        <otriple>Quetta_Development_Authority | headquarter | Quetta</otriple>
        <otriple>Daily_Awam | headquarter | Quetta</otriple>
        <otriple>Quetta_Electric_Supply_Company | locationCity | Quetta</otriple>
        <otriple>Hanna_Lake | nearestCity | Quetta</otriple>
        <otriple>Hazarganji-Chiltan_National_Park | nearestCity | Quetta</otriple>
        <otriple>Women_Chamber_of_Commerce_Quetta | regionServed | Quetta</otriple>
        <otriple>Safdar_Kiyani | residence | Quetta</otriple>
        <otriple>Safeer_Ullah_Khan | residence | Quetta</otriple>
        <otriple>Meena_Hazara | residence | Quetta</otriple>
      </originaltripleset>
    </entry>
  </entries>
</benchmark>
```


Figure 2: A long input for Quetta (City, size = 39 triples).

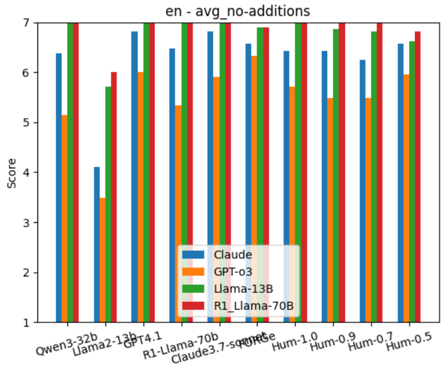

Figure 3: English, No-additions detailed results

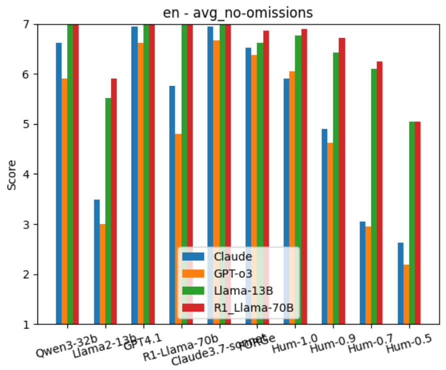

Figure 4: English, No-omissions detailed results

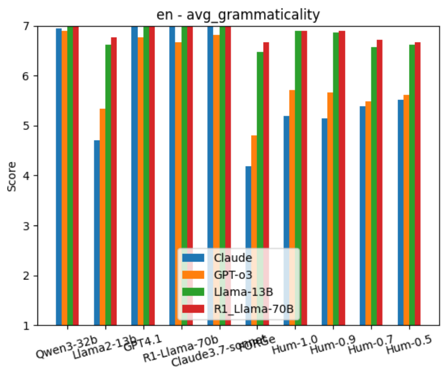

Figure 5: English, Grammaticality detailed results

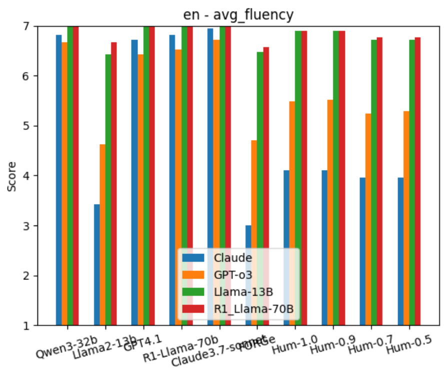

Figure 6: English, Fluency detailed results.

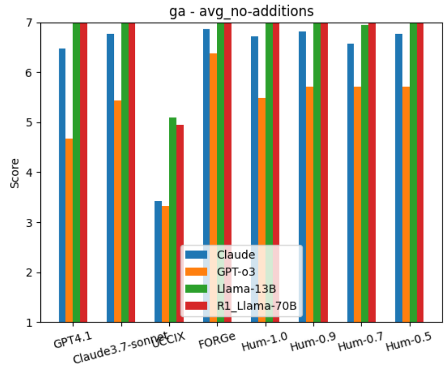

Figure 7: Irish, No-additions detailed results.

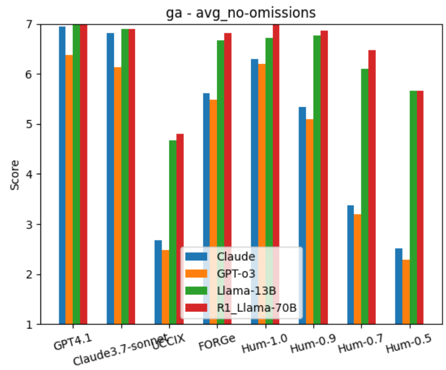

Figure 8: Irish, No-omissions detailed results.

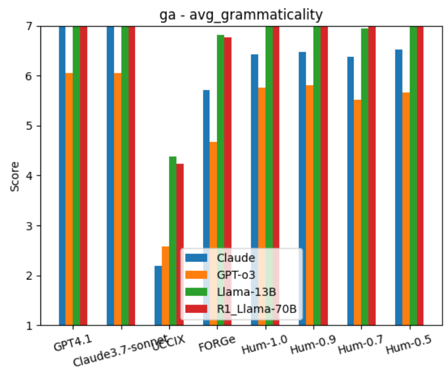

Figure 9: Irish, Grammaticality detailed results.

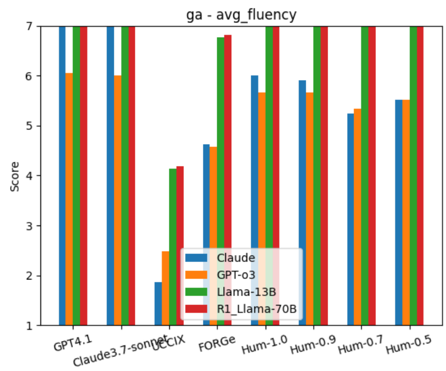

Figure 10: Irish, Fluency detailed results.

Prompt Name	Prompt Text
ENGLISH PROMPT	You are a data-to-text generation agent that transforms structured data in the form of subject–predicate–object (SPO) triples into fluent, informative, and human-like natural language text. **Your Goal:** Generate well-written paragraph(s) that convey all facts encoded in the triples while maintaining coherence and naturalness, as if written by a skilled human author. Your output should resemble a short article, report, or description — not a mechanical list of facts. **Process and Generation Guidelines** 1. **Analyze the Data:** • Identify distinct entities (subjects) and their associated facts. • Recognize relationships between entities to create narrative flow. • Group related information for logical organization. 2. **Plan Your Structure:** • Organize information in a natural, readable sequence—do not follow the input triple order rigidly. • Organize the text into coherent sentences and well-structured paragraphs, with each paragraph focusing on a specific topic or entity. • Group related entities and facts together to create coherent paragraphs (e.g., places, objects, biographical details, achievements, relationships, etc.). • Use paragraphs to separate distinct topics or entities, ensuring each paragraph has a clear focus. 3. **Write with Fluency and Variety:** • Use pronouns and natural references to avoid repetitive entity names. 4. **Ensure Complete Accuracy:** • Include every fact encoded in the triples without exception. • Never add external information or make inferences beyond the given data. • Preserve all factual content while using natural paraphrasing. • Cross-check that no information has been omitted from your final text. 5. **Maintain Professional Style:** • Write in third person with a neutral, encyclopedic tone. • Ensure grammatical correctness and proper punctuation. • Avoid bullet points, lists, or structured formatting. **What to Avoid** • Copying triples verbatim into the text. • Omitting any information from the triples. • Adding information not present in the triples. • Creating one sentence per triple (mechanical approach). • Using structured formats (XML, JSON, lists) instead of prose. • Generate only one prose using the data. Multiple prose is not allowed. **Output Requirements** • Return only the final generated text as continuous, fluent paragraph(s). Use multiple paragraphs when it improves organization and readability.
IRISH PROMPT	You are a data-to-text generation agent tasked with generating **natural, fluent Irish text** from structured data presented as subject–predicate–object triples written in **English**. **Task Objective:** Your goal is to verbalize all the information contained in the input triples in **authentic Irish**, producing a well-structured and human-like description or paragraph. The output should sound like it was written by a native Irish speaker, not a literal translation or a mechanical list of facts. **Input Format** • You will receive a list of RDF-style triples in **English**, for example: – (Person, birthDate, 1974) – (Person, occupation, “writer”) – (Writer, notableWork, “Book Title”) **Generation Guidelines** 1. **Comprehensive Coverage:** Use all facts presented in the triples. Do not omit or invent information. 2. **Linguistic Fluency:** Write in correct and idiomatic Irish. Use proper grammar, syntax, and vocabulary appropriate for formal writing or encyclopedic entries. 3. **Coherence & Flow:** Organize the facts into a natural narrative. Group related information into sentences and paragraphs. Avoid simply listing the facts in order. 4. **Cultural Appropriateness:** Adapt English names, locations, and conventions where needed to fit Irish usage or orthography (e.g., use Irish forms of countries, months, occupations if available). 5. **Avoid Literal Translation:** Do not translate the triples directly or word-for-word. Instead, reformulate them naturally in Irish. **Output Format** • Write only the Irish text. Do not include explanations, metadata, or translations of the triples. **Example Input** Triples: • (Douglas Hyde, birthPlace, Castlerea) • (Douglas Hyde, birthDate, 1860) • (Douglas Hyde, positionHeld, President of Ireland) **Example Output** Rugadh Dubhghlas de hÍde i gCaisleán Riabhach sa bhliain 1860. Bhí sé ina chéad Uachtarán ar Éirinn. Begin generating the Irish text now based on the input triples. [GENERATED TEXT]
INPUT PROMPT	Here are the subject–predicate–object triples to convert: `{triples}` Transform this structured data into coherent, flowing prose that naturally integrates all the factual information. Ensure every fact from the triples is represented in your text while maintaining readability and logical flow. [GENERATED TEXT]

Table 2: Generation prompts for English and Irish data-to-text realization tasks.