INLG 2025

Proceedings
of the
18th International Natural Language Generation Conference

Generation Challenges

October 29 – November 2, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-324-1

Preface

The Generation Challenges (GenChal) aim at bringing together a variety of shared-task efforts that involve the generation and evaluation of systems that produce natural language. This year again, the Generation Challenges will be held during a special session at the 18th International Conference on Natural Language Generation (INLG, October 29th- November 2nd 2025). The session traditionally comprises oral presentations of proposals for new shared tasks, and oral presentations of results by the organisers of recently completed tasks, while completed task participants present their submissions during a main conference poster session.

In 2025, the oral session will take place on Nov. 2nd at 14:00-15:30 Hanoi time; we have **one new task proposal** (Live Commentary Planning and Generation) and **one report on the human evaluation results of a completed task** (GEM English and Spanish Data-to-text), whose metrics results were presented in 2024. The 2025 oral session also includes one **task update** of a challenge proposed in 2024 (Long-Form Analogy Evaluation Challenge) and an **invited talk** to showcase the results of a recent large-scale shared task held outside of the GenChal umbrella (the ReproNLP shared task).

New challenge proposal (*in proceedings*):

- *Live Commentary Planning and Generation*
  Chung-Chi Chen, Ming-Hung Wang, Ramon Ruiz-Dolz, Chris Reed, Ichiro Kobayashi, Yusuke Miyao and Hiroya Takamura

  The proposal was reviewed positively by the three program committee members, who also provided valuable feedback to the task organisers.

Completed challenge overview (*in proceedings*):

- *The 2024 GEM Shared Task on Multilingual Data-to-Text Generation: English and Spanish Qualitative Evaluation Results*
  João Sedoc, Simon Mille, Miruna Adriana Clinciu, Yixin Liu, Kaustubh Dhole, Saad Mahamood

  The report comes along with two updated GEM participating system descriptions which include an analysis of the results of the released human evaluation (*in proceedings*), and will be presented during INLG's afternoon poster session of Nov. 1st.

Challenge update:

- *Long-Form Analogy Evaluation Challenge*
  Bhavya Bhavya, Chris Palaguachi, Yang Zhou, Suma Bhat, ChengXiang Zhai

  The challenge was proposed in 2024 and is still running at the time of the conference.

Invited talk:

- *ReproNLP Shared Task Overview*
  Anya Belz, Craig Thomson, Javier González Corbelle, Malo Ruelle

I would like to express my gratitude to the reviewers, the task organisers, as well as the INLG Programme Chairs, Publication Chair and Local Organisers for their precious help during the organisation process.

Simon Mille

This event is sponsored by Vingroup Innovation Foundation (VINIF – VinBigData).

Organizing Committee

Generation Challenge Chair

Simon Mille, ADAPT Research Centre / Dublin City University

Publication Chair

Ondřej Dušek, Charles University

Program Committee

David M. Howcroft, University of Aberdeen
Guy Lapalme, Université de Montréal
Anastasia Shimorina, Orange

Table of Contents

The 2024 GEM Shared Task on Multilingual Data-to-Text Generation: English and Spanish Qualitative Evaluation Results

João Sedoc[1], Simon Mille[2], Miruna Clinciu[3], Yixin Liu[4], Kaustubh Dhole[5], Saad Mahamood[6]

[1]New York University, [2]ADAPT, Dublin City University, [3]Heriot Watt University, [4]Yale University, [5]Emory University, [6]Shopware

Correspondence: jsedoc@stern.nyu.edu , simon.mille@adaptcentre.ie

Abstract

We present in this paper the results of the 2024 GEM shared task of multilingual data-to-text generation for both English and Spanish. In particular, we focus on evaluating the submitted systems across different datasets, metrics, and compare the results generated using LLM and human evaluators when given the same evaluation instructions. The results presented show that submitted systems that use more resources perform better and that while LLMs and humans are usually aligned in how they rank systems, the LLMs tend to award higher scores than humans. We describe the motivation for this shared task, describe the tasks and submitted systems, the evaluation setup, and the results obtained.

1 Introduction

The Generation, Evaluation, and Metrics (GEM) initiative (Gehrmann et al., 2021) has focused over the past four years on better comprehending and measuring the progress that the field of Natural Language Generation (NLG) has made, through the iterative creation of datasets (Mille et al., 2021), tools (Dhole et al., 2023), and the assessments of different text generation systems using human and/or automatic evaluation approaches (Gehrmann et al., 2022; Zhang et al., 2023; Nawrath et al., 2024). The focus on evaluation and its practices within NLG has enabled a better understanding of the current challenges that are present when evaluating such systems.

Given the broad adoption of (very) large language models (LLMs) within the field of NLG for both the process of generation and evaluation, it is important to better quantify and qualify the performance of LLMs against human evaluators on different tasks, so as to have a broader understanding of their strengths and weaknesses when used as a tool for either content creation or evaluation. This is especially important given the indications in research literature that LLMs may favour their own output (Panickssery et al., 2024), have issues with respect to data contamination (Balloccu et al., 2024), suffer from biases (Kotek et al., 2023), inconsistencies (Dhole et al., 2025), hallucination, lack of semantic faithfulness (Gehrmann et al., 2023), etc. Nevertheless, there has been significant interest in exploring the use of LLMs for the task of evaluation in NLG (Gao et al., 2025), largely driven by the considerable challenges met when conducting human evaluations. Difficulties in recruiting high-quality annotators (Zhang et al., 2023), lack of robust evaluation methodology (Thomson and Reiter, 2020) and poor reporting practices (Howcroft et al., 2020) for instance have made human evaluations difficult to run, interpret and compare with one another. With this shared task report, we aim to analyse the performance of LLMs both as content generators and as quality evaluators, by looking at multiple aspects such as datasets with different properties (e.g. in-domain, out-of-domain), different languages (English and Spanish), different evaluation criteria (e.g. `Fluency` and `Grammaticality`), etc. across multiple submitted systems from task participants.

In the future, we will follow up with a report presenting results obtained for the Swahili language, in both the data-tot-text and summarisation tasks, since unfortunately the Swahili evaluations are still running at the time this paper is being published.

In this paper we summarise the GEM 2024 data-to-text generation tasks and the participating systems (Section 2), we describe the qualitative evaluation setup by detailing the data, LLM and human evaluation approaches (Section 3), and we present the results for the data-to-text task with multiple sets of analyses, including a discussion of the instance-level and system-level correlations (Section 4). In the final section (Section 5) we discuss our conclusions and the main findings from this shared task.

Proceedings of the 18th International Natural Language Generation Conference: Generation Challenges, pages 1–36
October 29–November 2, 2025.

Team	D2T-1	D2T-2	Implementation
DCU-ADAPT-modPB (Osuji et al., 2024)	en, sw		Flan-T5-0.7B (FT) + GPT-4 + MT
DCU-NLG-PBN (Lorandi and Belz, 2024)	en, es, sw	en, es, sw	Mistral-7B-Instruct (FT) + MT
DCU-NLG-Small (Mille et al., 2024a)	en, es, sw	en, es, sw	Rules + T5-0.2B (FT) + MT
DipInfo-UniTo (Oliverio et al., 2024)	en	en	Rules + Mistral-7B (FT) + Mistral-7B
OSU CompLing (Allen et al., 2024)	en, es	en, es	One Llama2-7B (FT) per language
RDFpyrealb (Lapalme, 2024)	en	en	Rules
SaarLST (Jobanputra and Demberg, 2024)	en	en	Rules + Mixtral-8x7B (RAG)
Anonymous (withdrawn)	en	en	N/A

Table 1: Overview of evaluated systems; en=English, es=Spanish, sw=Swahili. "+" means that the components on each side are pipelined; "FT" = "fine-tuned"; "MT" = "machine translation" (when generating in other languages than English); "RAG" = "Retrieval-Augmented Generation". For more details see the respective papers.

2 Summary of tasks and participating systems

In this section, we provide a brief overview of the tasks and participants; for more details, see (Mille et al., 2024b). The GEM 2024 data-to-text task consisted of generating texts starting from small knowledge graphs of between 2 and 7 triples. It had 2 subtasks, one that uses DBpedia triples (D2T-1), as in the WebNLG shared task (Gardent et al., 2017), the other one that uses newly collected Wikidata triples (D2T-2). For each subtask, 3 versions of the same inputs were provided, as described in (Axelsson and Skantze, 2023): a factual (FA) version, with factually correct data (e.g. *Barack_Obama, birthYear, 1961*) a counterfactual (CFA) version, in which entities were swapped with other entities of the same category (e.g. *Lady_Gaga, birthYear, 1961*), and a fictional version (FI), in which entities and values were created using an LLM (e.g. *Wonyer_Lator, birthYear, 4397*). Participants submitted outputs in up to 9 languages. The participating teams and details of their submissions used in the qualitative evaluation are provided in Table 1.

3 Qualitative evaluation setup

The system outputs, code for running evaluations and computing results, plots and correlations are publicly available on GitHub `https://github.com/GEM-benchmark/human-eval-shared-task-2024`.[1]

3.1 Evaluated data

For the data-to-text task, we evaluated all outputs in English, Spanish and Swahili as shown in Table 1. Every time a language appears in column D2T-1 or D2T-2, it means that 3 datasets (FA, CFA, FI, see Section 2) of 180 input/output pairs were evaluated.

	# Systems		# Input/output pairs	
Dataset↓	en	es	en	es
D2T-1-FA	8+1	3	1,620	540
D2T-1-CFA	8	3	1,440	540
D2T-1-FI	8	3	1,440	540
D2T-2-FA	7	3	1,260	540
D2T-2-CFA	7	3	1,260	540
D2T-2-FI	7	3	1,260	540
TOTAL			8,280	3,240

Table 2: Number of evaluated data points for the data-to-text task. Each dataset has 180 data points. +1 on the D2T-1-FA row is the human-written WebNLG texts.

For the D2T-1-FA subset, we also evaluated 180 original WebNLG 2020 human-written texts, by selecting a random text for each of the 180 sampled data points. The total number of input/output pairs evaluated is thus 8,280 in English, 3,240 in Spanish, and 2,700 in Swahili;[2] we show the breakdown of the count for English and Spanish in Table 2. Note that in Section 4, there is one less system than the number of evaluated systems for English; this is simply because one team withdrew their submission, so we do not report their evaluation results in this paper. Also, since DCU-ADAPT-modPB did not submit outputs for the D2T-2 subtask, there is a different number of system outputs between D2T-1 and D2T-2 for English (but not for Spanish because they did not submit outputs in this language).

3.2 Human evaluation

In (Mille et al., 2024b), we provide details on the evaluator recruitment and training processes, and the evaluation criteria and task design. Table 3 is replicated here to detail the four dimensions used for data-to-text. We refer to `No-Omissions` and `No-Additions` as "semantic accuracy criteria", since they both assess to what extent the semantic contents of the outputs match those of the inputs.

[1] One half of the human evaluation annotations will have a delayed release so as to keep an undisclosed set of results.

[2] The Swahili evaluation is still running.

Criterion name	Definition
No-Omissions	ALL the information in the table is present in the text.
No-Additions	ONLY information from the table 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.

Table 3: Criteria used for data-to-text generation

On the other hand, `Grammaticality` and `Fluency` assess qualities of the output texts in their own right, regardless of the input; below, we refer to these together as the "intrinsic quality criteria".

All input/output pairs were assessed by at least 2 human evaluators, and a subset of the data was annotated by several evaluators to carry out inter-annotator agreement (IAA) analysis:

- For English, the IAA subset consists of 2 outputs per system and per dataset, for a total of 80 input/output pairs, all scored by the same 18 evaluators.
- For Spanish, the IAA subset consists of 4 outputs per system and per dataset, for a total of 72 input/output pairs, all scored by the same 7 evaluators.

For English, we collected a total of 62,756 individual ratings (15,689 rating per quality criterion), and for Spanish 27,936 individual ratings (6,984 ratings per quality criterion).

For the human evaluation results, we computed the mean score across all individual scores received by each system on each dataset. To indicate whether the average scores between two systems is significant, we computed the system ranking using Tukey's Honestly Significant Difference (Tukey's HSD; Tukey, 1949), which tests all pairwise differences between systems while correcting for multiple comparisons. This measure has the advantage of allowing partial ties between systems. We set our threshold for statistical significance to 0.05.

We calculated the inter-annotator agreement between annotators both on the high overlap dataset and the result of the dataset using Krippendorff's alpha (Krippendorff, 1970). This is a commonly used annotation method when not all annotators annotate all items. Given our 7-point Likert scale, we use interval metric penalization rather than nominal or linear.

All annotators were recruited on Amazon Mechanical Turk. We follow the filters from Zhang et al. (2023). For the English task location was limited to US; however, for the Spanish task there was no location limit and instead a language fluency requirement of both English and Spanish. We require our workers to have a minimum of 1,000 completed tasks and 97% completion rate. All annotators were required to pass a training and filtering task. Annotators were further filtered out during on the inter-annotator agreement subset of our dataset where workers with an average Cohen's Kappa under 0.5 were removed from further annotations. Annotators were paid on a task basis with an expected compensation of roughly $15 per hour.

3.3 LLM-as-Judge evaluation

We evaluated all English and Spanish outputs detailed in Table 2 according to all four quality criteria. We chose four LLMs for their complementary strengths: **o3-mini**[3][4] (OpenAI, 2024) via the OpenAI API,[5] for its improved performance on reasoning benchmarks, **DeepSeek-R1-Distill-Llama-70B** (DeepSeek-AI, 2025) to provide a locally reproducible open-weight baseline, and **Gemini-1.5-Flash**[6] (Gemini Team, 2024) via the aiXplain API[7] and **GPT-4o-mini**[8] (OpenAI, 2024) for their speed, cost-effectiveness, and broad accessibility. 4 Nvidia A6000 GPUs were used for hosting the DeepSeek model.

The total cost on the aiXplain and OpenAI platforms was below $60 (∼$6 for Gemini-1.5-Flash, ∼$2 for GPT-4o-mini, and ∼$50 for o3-mini).

The prompts sent to the LLMs contained the same information as provided to the human annotators; see a sample prompt in Appendix A. The Notebooks used to run the aiXplain and OpenAI evaluations can be found on GitHub.[9]

For 3 English input/output pairs, DeepSeek-R1-Distill-Llama-70B did not return any score; As a result, in English, three systems have 179 instead of 180 sets of four scores (one score per criterion) for the D2T-2-FI dataset (DCU-NLG-Small,

[3] https://openai.com/index/openai-o3-mini
[4] Model ID *o3-mini-2025-01-31*.
[5] https://openai.com/api/
[6] Model ID: *674b73f06eb563a748561d41*
[7] https://platform.aixplain.com/dashboard/
[8] Model ID *gpt-4o-mini-2024-07-18*.
[9] https://github.com/mille-s/GEM24_EvalLLM

DipInfo-UniTo and SaarLST). For the other five datasets, all systems have all scores. In total, we collected ratings (i) for 8,280 English and 3240 Spanish input/output pairs, (ii) with four different LLMs and (iii) for four criteria, for a total of (((8,280+3,240)*4)-3)*4 = 184,308 individual ratings.

In the results section, we report averaged LLM scores, obtained by computing the mean of the four LLMS for each evaluated instance (and then the mean over all considered instance, using Tukey's HSD for ranks, see Section 3.2); Appendix C shows all individual LLM scores for each system on each dataset.

3.4 Computation of correlations between human and LLM-as-judge evaluations and between evaluation dimensions

Scores for all systems on all datasets were aligned in different CSV files. From each filename we extract a *slice* (system, subset, evaluator) and define a shared row key $u = (\text{id}, \text{system}, \text{subset\_eval})$, ensuring that all models evaluated on the same slice use identical u for the same item. Stacking all files and pivoting yields a wide matrix $W \in \mathbb{R}^{n \times p}$ whose rows index aligned items u and whose columns are features $f = (d, m)$ formed by a dimension d (`No-Omissions`, `No-Additions`, `Grammaticality`, `Fluency`) and a model m. The entry $W_{i,f}$ is the score for item i under feature f.

For each pair of feature columns a and b, correlations are computed on the pairwise-complete set $S_{ab} = \{\, i : W_{i,a} \text{ and } W_{i,b} \text{ are both observed} \,\}$ with size $N_{ab} = |S_{ab}|$. Let $R_a(i)$ and $R_b(i)$ denote the midranks of $W_{i,a}$ and $W_{i,b}$ over $i \in S_{ab}$.

To address multiple testing, we control the false discovery rate across the unique pairs using the Benjamini–Hochberg procedure. If $p_{(1)} \le \cdots \le p_{(m)}$ are the ordered p-values for the m tests, the corresponding q-values are

$$q_{(k)} = \min_{j \ge k} \frac{m}{j}\, p_{(j)}, \qquad k = 1, \ldots, m,$$

which are mapped back to their original pairs to yield q_{ab}. Cells are annotated with ρ_{ab} and significance stars for $q_{ab} < 0.05, 0.01, 0.001$ (shown as $*, **, ***$). Diagonal entries satisfy $\rho_{aa} = 1$ with $p_{aa} = q_{aa} = 0$, and the displayed N_{ab} is the pairwise-complete sample size used for each ρ_{ab}. The resulting heatmaps of Section 4.2 visualize ρ_{ab} on a fixed $[-1, 1]$ diverging scale. More details are provided in Appendix D.

4 Data-to-text qualitative evaluation results

In this section, we present (i) a detailed analysis of the human and LLM-as-judge results for all datasets and all criteria for both languages (Section 4.1), (ii) an analysis of the instance-level correlations between human and LLM ratings (Section 4.2), (iii) an analysis of system-level correlations between human ratings, LLM ratings and metrics (section 4.3), and (iv) an analysis of human annotator behaviour (Section 4.4).

4.1 Evaluation results

Figures 1 to 7 show the system rankings for English outputs, and Figures 8 to 13 show the system rankings for Spanish outputs. In all figures, the left-hand side tables report on human evaluations, while the right-hand side tables report on the LLM-as-judge evaluation. Each side of the figures comprises four table, which correspond to the four evaluated quality criteria (in descending order: `No-Omissions`, `No-Additions`, `Grammaticality`, `Fluency`). Each table row contains a system along with its mean score for the given criterion on the given dataset, and groupings based on Tukey's HSD post-hoc test, which denote statistically significant differences between systems (i.e. the scores of two systems who share a letter in the same table do not have statistically significant differences). The mean human evaluation scores were computed with two or more scores on 180 data points per system, while the mean LLM evaluation scores we computed with four different LLMs on 180 data points per system (see Section 3).

For each language, we report the following:

- System rankings for each criterion on in-domain data only (D2T-1-FA, D2T-1-CFA, D2T-1-FI); Figures 1 and 8.
- System rankings for each criterion on out-of-domain data only (D2T-2-FA, D2T-2-CFA, D2T-2-FI); Figures 2 and 9.
- System rankings for each criterion on factual data only (D2T-1-FA, D2T-2-FA); Figures 3 and 10.
- System rankings for each criterion on counterfactual data only (D2T-1-CFA, D2T-2-CFA); Figures 4 and 11.

(a) No-omissions human-en-D2T-1-*

	Mean	Group
SaarLST	5.75	A
RDFpyrealb	5.72	A
DipInfo-UniTo	5.48	B
DCU-NLG-PBN	5.44	B
DCU-ADAPT-modPB	5.33	B
OSU-CompLing	4.77	C
DCU-NLG-Small	4.59	D

(b) No-omissions llm-en-D2T-1-*

	Mean	Group
RDFpyrealb	6.81	A
SaarLST	6.78	A
DipInfo-UniTo	6.55	B
DCU-NLG-PBN	6.54	B
OSU-CompLing	6.18	C
DCU-ADAPT-modPB	6.16	C
DCU-NLG-Small	5.71	D

(c) No-additions human-en-D2T-1-*

	Mean	Group
DCU-ADAPT-modPB	5.62	A
SaarLST	5.50	AB
DipInfo-UniTo	5.48	AB
RDFpyrealb	5.47	AB
DCU-NLG-PBN	5.38	B
OSU-CompLing	4.64	C
DCU-NLG-Small	4.50	C

(d) No-additions llm-en-D2T-1-*

	Mean	Group
DCU-ADAPT-modPB	6.91	A
DipInfo-UniTo	6.84	AB
RDFpyrealb	6.82	BC
DCU-NLG-PBN	6.79	BC
SaarLST	6.75	C
OSU-CompLing	6.62	D
DCU-NLG-Small	6.26	E

(e) Grammaticality human-en-D2T-1-*

	Mean	Group
DCU-ADAPT-modPB	6.21	A
SaarLST	5.96	B
DCU-NLG-PBN	5.88	B
DipInfo-UniTo	5.83	B
DCU-NLG-Small	5.30	C
OSU-CompLing	5.21	C
RDFpyrealb	4.69	D

(f) Grammaticality llm-en-D2T-1-*

	Mean	Group
DCU-ADAPT-modPB	6.98	A
DipInfo-UniTo	6.96	A
SaarLST	6.96	A
DCU-NLG-PBN	6.94	A
DCU-NLG-Small	6.84	B
OSU-CompLing	6.74	C
RDFpyrealb	6.14	D

(g) Fluency human-en-D2T-1-*

	Mean	Group
DCU-ADAPT-modPB	6.12	A
SaarLST	5.89	B
DCU-NLG-PBN	5.82	BC
DipInfo-UniTo	5.73	C
OSU-CompLing	5.29	D
DCU-NLG-Small	5.25	D
RDFpyrealb	4.81	E

(h) Fluency llm-en-D2T-1-*

	Mean	Group
DCU-ADAPT-modPB	6.97	A
SaarLST	6.93	AB
DCU-NLG-PBN	6.92	AB
DipInfo-UniTo	6.91	B
OSU-CompLing	6.81	C
DCU-NLG-Small	6.73	D
RDFpyrealb	6.11	E

Figure 1: System rankings for English in-domain data (left: human ratings, right: llm ratings)

- System rankings for each criterion on fictional data only (D2T-1-FI, D2T-2-FI); Figures 5 and 12.

- System rankings for each criterion on all datasets (D2T-1-FA, D2T-1-CFA, D2T-1-FI, D2T-2-FA, D2T-2-CFA, D2T-2-FI); Figures 7 and 13.

Additionally, for English, we report system rankings on in-domain factual data only (D2T-1-FA, Figure 6), since only for this dataset we were able to evaluate human-written texts. Note that for each table we only consider systems that have outputs for all selected datasets.

All scores on individual datasets for English and Spanish can be found in Appendix B.

4.1.1 Results for English in-domain and out-of-domain data

Figures 1 and 2 show the rankings on the English in-domain and out-of-domain data respectively.

In-domain: Semantic accuracy criteria. For humans, SaarLST and RDFpyrealb are above others, while DCU-NLG-Small gets the lowest scores. The lowest-ranking systems are the same in all four tables (OSU-CompLing, followed by DCU-NLG-Small). For `No-Omissions`, the system rankings and groupings are very similar in the human and the LLM tables; only DCU-ADAPT-modPB is one group lower in the LLM rankings. For `No-Additions`, rankings and groupings are again similar in human and LLM tables, with the notable exception of SaarLST which is rank quite lower by LLM judges than by humans.

(a) No-omissions human-en-D2T-2-*

	Mean	Group
SaarLST	6.03	A
DipInfo-UniTo	5.69	B
DCU-NLG-PBN	5.42	C
RDFpyrealb	5.42	C
OSU-CompLing	4.69	D
DCU-NLG-Small	4.34	E

(b) No-omissions llm-en-D2T-2-*

	Mean	Group
SaarLST	6.92	A
RDFpyrealb	6.74	B
DipInfo-UniTo	6.60	C
DCU-NLG-PBN	6.57	C
OSU-CompLing	6.20	D
DCU-NLG-Small	5.20	E

(c) No-additions human-en-D2T-2-*

	Mean	Group
SaarLST	5.81	A
DipInfo-UniTo	5.72	A
DCU-NLG-PBN	5.24	B
RDFpyrealb	5.00	C
OSU-CompLing	4.49	D
DCU-NLG-Small	4.12	E

(d) No-additions llm-en-D2T-2-*

	Mean	Group
SaarLST	6.86	A
DipInfo-UniTo	6.83	A
DCU-NLG-PBN	6.68	B
RDFpyrealb	6.65	BC
OSU-CompLing	6.57	C
DCU-NLG-Small	5.58	D

(e) Grammaticality human-en-D2T-2-*

	Mean	Group
SaarLST	6.12	A
DipInfo-UniTo	5.87	B
DCU-NLG-PBN	5.71	C
OSU-CompLing	5.09	D
DCU-NLG-Small	4.89	E
RDFpyrealb	4.18	F

(f) Grammaticality llm-en-D2T-2-*

	Mean	Group
SaarLST	6.98	A
DCU-NLG-PBN	6.96	A
DipInfo-UniTo	6.88	B
OSU-CompLing	6.76	C
DCU-NLG-Small	6.67	D
RDFpyrealb	5.62	E

(g) Fluency human-en-D2T-2-*

	Mean	Group
SaarLST	6.06	A
DipInfo-UniTo	5.83	B
DCU-NLG-PBN	5.67	C
OSU-CompLing	5.16	D
DCU-NLG-Small	4.89	E
RDFpyrealb	4.39	F

(h) Fluency llm-en-D2T-2-*

	Mean	Group
SaarLST	6.97	A
DCU-NLG-PBN	6.94	A
DipInfo-UniTo	6.84	B
OSU-CompLing	6.83	B
DCU-NLG-Small	6.49	C
RDFpyrealb	5.59	D

Figure 2: System rankings for English out-of-domain data (left: human ratings, right: llm ratings)

In-domain: Intrinsic quality criteria. Both human and LLM evaluations place DCU-ADAPT-modPB first for both `Grammaticality` and `Fluency`, and the same three systems last DCU-NLG-Small/OSU-CompLing followed by RDFpyrealb for humans, and DCU-NLG-Small followed by OSU-CompLing followed by RDFpyrealb for LLMs. DCU-ADAPT-modPB ranks first alone in the human tables, while LLMs have more difficulty distinguishing between the top four systems, which end up in one single or two groups.

Out-of-domain: Semantic accuracy criteria. In the human evaluation results, SaarLST, who used the second largest model after DCU-ADAPT-modPB, is ranked first for both criteria, and DipInfo-UniTo is in the same group only for `No-Additions`. DCU-NLG-PBN and RDFpyrealb obtain similar scores, even though the former ranks higher for `No-Additions`. As for the in-domain data, OSU-CompLing and DCU-NLG-Small rank at the bottom, but this time DCU-NLG-Small ranks lower for both criteria. LLMs situate RDFpyrealb one rank higher for both criteria.

Out-of-domain: Intrinsic quality criteria. For `Grammaticality` and `Fluency`, the picture is very clear in the human evaluation results with the same rankings and one system per group. The systems are in the same order as for the semantic accuracy criteria except for RDFpyrealb, which ranks last for both criteria, as it was the case on the in-domain data. The main difference between human and LLM-as-judge evaluations is that the rankings between DCU-NLG-PBN and DipInfo-UniTo, for which humans prefer the latter while LLMs prefer the former. Another difference is that LLMs produce several ties between teams, while humans did not have any, which may reflect the higher level of difficulty for LLMs in judging intrinsic text qualities.

(a) No-omissions human-en-*-FA

	Mean	Group
SaarLST	5.99	A
DipInfo-UniTo	5.63	B
RDFpyrealb	5.60	B
DCU-NLG-PBN	5.48	B
OSU-CompLing	4.99	C
DCU-NLG-Small	4.66	D

(b) No-omissions llm-en-*-FA

	Mean	Group
SaarLST	6.91	A
RDFpyrealb	6.80	A
DCU-NLG-PBN	6.58	B
DipInfo-UniTo	6.58	B
OSU-CompLing	6.31	C
DCU-NLG-Small	5.58	D

(c) No-additions human-en-*-FA

	Mean	Group
SaarLST	5.88	A
DipInfo-UniTo	5.82	A
DCU-NLG-PBN	5.52	B
RDFpyrealb	5.27	C
OSU-CompLing	4.91	D
DCU-NLG-Small	4.63	E

(d) No-additions llm-en-*-FA

	Mean	Group
DipInfo-UniTo	6.88	A
SaarLST	6.86	A
DCU-NLG-PBN	6.79	AB
RDFpyrealb	6.73	B
OSU-CompLing	6.69	B
DCU-NLG-Small	5.96	C

(e) Grammaticality human-en-*-FA

	Mean	Group
SaarLST	6.18	A
DipInfo-UniTo	6.07	A
DCU-NLG-PBN	6.06	A
OSU-CompLing	5.54	B
DCU-NLG-Small	5.26	C
RDFpyrealb	4.32	D

(f) Grammaticality llm-en-*-FA

	Mean	Group
DCU-NLG-PBN	6.96	A
SaarLST	6.96	A
DipInfo-UniTo	6.93	A
OSU-CompLing	6.79	B
DCU-NLG-Small	6.73	B
RDFpyrealb	5.76	C

(g) Fluency human-en-*-FA

	Mean	Group
SaarLST	6.11	A
DCU-NLG-PBN	5.98	A
DipInfo-UniTo	5.97	A
OSU-CompLing	5.58	B
DCU-NLG-Small	5.24	C
RDFpyrealb	4.52	D

(h) Fluency llm-en-*-FA

	Mean	Group
SaarLST	6.96	A
DCU-NLG-PBN	6.95	A
DipInfo-UniTo	6.88	AB
OSU-CompLing	6.86	B
DCU-NLG-Small	6.61	C
RDFpyrealb	5.72	D

Figure 3: System rankings for English factual data (left: human ratings, right: llm ratings)

Comparison between in-domain and out-of-domain data. For in-domain data, DCU-ADAPT-modPB, the only submission that used a very large language model (GPT-4), obtained the best scores on three out of four criteria, with an apparent issue with omitting parts of the input (lower `No-Omissions` rankings). DCU-ADAPT-modPB did not submit outputs for the out-of-domain data but all other teams did. The two systems that use a rule-based component as main generation engine (RDFpyrealb and DCU-NLG-Small) see their scores clearly drop on out-of-domain data, while for the other systems the scores are rather similar. DipInfo-UniTo and especially SaarLST even obtain higher scores on the out-of-domain data. We observe that there are considerably less ties on the out-of-domain data, which possibly reflects the fact that system outputs are less homogenous on this dataset. Further statistical testing is necessary to test if this is indeed significant.

4.1.2 Results for English factual, counterfactual and fictional data

Figures 3, 4 and 5 show the rankings on the English factual, counterfactual and fictional data respectively.

Factual: Semantic accuracy criteria. In the human evaluation results, SaarLST ranks first for both criteria, along with DipInfo-UniTo for `No-Additions`. DCU-NLG-PBN ranks in the second group for both criteria, while RDFpyrealB ranks second and third on `No-Omissions` and `No-Additions` respectively (as expected because of the lower scores of this system on out-of-domain data). OSU-CompLing and DCU-NLG-Small rank at the bottom, in this order. As observed for the out-of-domain data above, RDFpyrealB is positioned one rank higher by LLMs, which otherwise provide very similar rankings to humans, but once again with more ties between systems.

(a) No-omissions human-en-*-CFA

	Mean	Group
SaarLST	5.72	A
DipInfo-UniTo	5.58	A
RDFpyrealb	5.57	A
DCU-NLG-PBN	5.33	B
OSU-CompLing	4.74	C
DCU-NLG-Small	4.36	D

(b) No-omissions llm-en-*-CFA

	Mean	Group
SaarLST	6.76	A
RDFpyrealb	6.71	A
DipInfo-UniTo	6.51	B
DCU-NLG-PBN	6.40	B
OSU-CompLing	6.14	C
DCU-NLG-Small	5.28	D

(c) No-additions human-en-*-CFA

	Mean	Group
DipInfo-UniTo	5.54	A
SaarLST	5.33	AB
RDFpyrealb	5.19	BC
DCU-NLG-PBN	5.09	C
OSU-CompLing	4.49	D
DCU-NLG-Small	4.06	E

(d) No-additions llm-en-*-CFA

	Mean	Group
DipInfo-UniTo	6.80	A
RDFpyrealb	6.68	AB
SaarLST	6.66	B
DCU-NLG-PBN	6.60	B
OSU-CompLing	6.58	B
DCU-NLG-Small	5.79	C

(e) Grammaticality human-en-*-CFA

	Mean	Group
SaarLST	5.95	A
DipInfo-UniTo	5.81	AB
DCU-NLG-PBN	5.67	B
OSU-CompLing	4.97	C
DCU-NLG-Small	4.91	C
RDFpyrealb	4.38	D

(f) Grammaticality llm-en-*-CFA

	Mean	Group
SaarLST	6.96	A
DCU-NLG-PBN	6.93	A
DipInfo-UniTo	6.92	A
DCU-NLG-Small	6.70	B
OSU-CompLing	6.66	B
RDFpyrealb	5.75	C

(g) Fluency human-en-*-CFA

	Mean	Group
SaarLST	5.88	A
DipInfo-UniTo	5.74	AB
DCU-NLG-PBN	5.62	B
OSU-CompLing	5.06	C
DCU-NLG-Small	4.87	D
RDFpyrealb	4.52	E

(h) Fluency llm-en-*-CFA

	Mean	Group
SaarLST	6.93	A
DCU-NLG-PBN	6.90	A
DipInfo-UniTo	6.87	A
OSU-CompLing	6.74	B
DCU-NLG-Small	6.53	C
RDFpyrealb	5.70	D

Figure 4: System rankings for English counterfactual data (left: human ratings, right: llm ratings)

Factual: Intrinsic quality criteria. Humans prefer SaarLST, DipInfo-UniTo and DCU-NLG-PBN, all in the first group for both criteria, and place OSU-CompLing, DCU-NLG-Small and RDFpyrealb in the second third and fourth groups respectively. LLMs rank OSU-CompLing and DCU-NLG-Small higher for `Fluency` and `Grammaticality` respectively.

Counterfactual: Semantic accuracy criteria. For counterfactual data, SaarLST and DipInfo-UniTo are in the first group of the semantic accuracy tables, along with RDFpyrealb for `No-Omissions`. DCU-NLG-PBN, OSU-CompLing and DCU-NLG-Small then rank in this order. LLMs place DipInfo-UniTo one rank lower for `No-Omissions`, SaarLST one rank lower for `No-Additions`, and RDFpyrealb one rank above (first group) for `No-Additions`.

Counterfactual: Intrinsic quality criteria. For `Grammaticality` and `Fluency`, SaarLST and DipInfo-UniTo are again at the top, while DCU-NLG-PBN is in the same group as DipInfo-UniTo (but not SaarLST). OSU-CompLing, DCU-NLG-Small and RDFpyrealb then rank in this order, except for `Grammaticality`, for which OSU-CompLing and DCU-NLG-Small are tied. DCU-NLG-PBN is ranked in the first group by LLMs (i.e. one rank higher for both criteria when compared to human rankings).

Fictional: Semantic accuracy criteria. For both criteria, SaarLST is the only system in the first group, DipInfo-UniTo, RDFpyrealb and DCU-NLG-PBN are in the second group and OSU-CompLing and DCU-NLG-Small are in the third group. The LLM groupings are different, with RDFpyrealb in the first group for `No-Omissions`, and DipInfo-UniTo, RDFpyrealb and DCU-NLG-PBN in the first group for `No-Additions`, while for both criteria OSU-CompLing and DCU-NLG-Small are in consecutive groups, OSU-CompLing ranking higher.

(a) No-omissions human-en-*-FI

	Mean	Group
SaarLST	5.95	A
DipInfo-UniTo	5.55	B
RDFpyrealb	5.54	B
DCU-NLG-PBN	5.48	B
OSU-CompLing	4.47	C
DCU-NLG-Small	4.38	C

(b) No-omissions llm-en-*-FI

	Mean	Group
SaarLST	6.88	A
RDFpyrealb	6.82	A
DCU-NLG-PBN	6.69	B
DipInfo-UniTo	6.63	B
OSU-CompLing	6.11	C
DCU-NLG-Small	5.51	D

(c) No-additions human-en-*-FI

	Mean	Group
SaarLST	5.76	A
DipInfo-UniTo	5.43	B
DCU-NLG-PBN	5.32	B
RDFpyrealb	5.25	B
OSU-CompLing	4.28	C
DCU-NLG-Small	4.24	C

(d) No-additions llm-en-*-FI

	Mean	Group
SaarLST	6.89	A
DCU-NLG-PBN	6.82	A
DipInfo-UniTo	6.82	A
RDFpyrealb	6.79	A
OSU-CompLing	6.53	B
DCU-NLG-Small	6.00	C

(e) Grammaticality human-en-*-FI

	Mean	Group
SaarLST	5.99	A
DipInfo-UniTo	5.68	B
DCU-NLG-PBN	5.65	B
DCU-NLG-Small	5.11	C
OSU-CompLing	4.94	D
RDFpyrealb	4.62	E

(f) Grammaticality llm-en-*-FI

	Mean	Group
SaarLST	6.98	A
DCU-NLG-PBN	6.95	A
DipInfo-UniTo	6.91	A
DCU-NLG-Small	6.83	B
OSU-CompLing	6.80	B
RDFpyrealb	6.15	C

(g) Fluency human-en-*-FI

	Mean	Group
SaarLST	5.94	A
DCU-NLG-PBN	5.64	B
DipInfo-UniTo	5.62	B
DCU-NLG-Small	5.09	C
OSU-CompLing	5.03	C
RDFpyrealb	4.77	D

(h) Fluency llm-en-*-FI

	Mean	Group
SaarLST	6.97	A
DCU-NLG-PBN	6.95	A
DipInfo-UniTo	6.87	B
OSU-CompLing	6.86	B
DCU-NLG-Small	6.68	C
RDFpyrealb	6.13	D

Figure 5: System rankings for English fictional data (left: human ratings, right: llm ratings)

Fictional: Intrinsic quality criteria. Here too, for both criteria, SaarLST is the only system in the first group, but only DipInfo-UniTo and DCU-NLG-PBN are in the second group, followed by OSU-CompLing and DCU-NLG-Small in this order, and RDFpyrealb at the bottom. DCU-NLG-PBN is positioned in the first group by LLMs for both criteria; unlike human evaluators, LLMs tie OSU-CompLing and DCU-NLG-Small for `Grammaticality` but ranks the former higher in terms of `Fluency`.

Comparison between factual, counterfactual and fictional data. In the human evaluation, across criteria, scores for all systems but RDFpyrealb are lower on the counterfactual and fictional datasets compared to the scores on the factual dataset. RDFpyrealb maintains almost all its scores on the counterfactual and fictional datasets, with even higher scores (although still under 5) for `Grammaticality` and `Fluency` on the fictional dataset. According to the LLM-as-judge evaluation, the score drop between the factual and counterfactual datasets is much less evident, in particular for the intrinsic quality criteria. When comparing factual and fictional dataset scores, LLMs essentially give the same scores as the respective human scores to all systems but RDFpyrealb, which gets higher scores especially for the intrinsic quality criteria. Human evaluations produce slightly more rank ties on the counterfactual and fictional datasets than they do on the factual dataset.

4.1.3 Results for English in-domain factual data

Figure 6 shows the rankings on the English in-domain factual data; this table is the only one that contains all system outputs along with human-written references from WebNLG 2020 (Castro Ferreira et al., 2020). Also note that the inputs and human-written references for this test set have been

(a) No-omissions human-en-D2T-1-FA

	Mean	Group
SaarLST	5.79	A
RDFpyrealb	5.74	AB
DCU-NLG-PBN	5.49	BC
DipInfo-UniTo	5.45	C
DCU-ADAPT-modPB	5.42	CD
WebNLG-Human	5.14	DE
OSU-CompLing	4.99	E
DCU-NLG-Small	4.88	E

(b) No-omissions llm-en-D2T-1-FA

	Mean	Group
RDFpyrealb	6.86	A
SaarLST	6.86	A
WebNLG-Human	6.67	AB
DCU-NLG-PBN	6.58	B
DipInfo-UniTo	6.51	BC
OSU-CompLing	6.32	CD
DCU-ADAPT-modPB	6.14	DE
DCU-NLG-Small	6.01	E

(c) No-additions human-en-D2T-1-FA

	Mean	Group
DCU-ADAPT-modPB	5.82	A
SaarLST	5.61	AB
DipInfo-UniTo	5.59	AB
DCU-NLG-PBN	5.56	AB
RDFpyrealb	5.41	B
WebNLG-Human	5.05	C
OSU-CompLing	4.85	C
DCU-NLG-Small	4.85	C

(d) No-additions llm-en-D2T-1-FA

	Mean	Group
DCU-ADAPT-modPB	6.95	A
DCU-NLG-PBN	6.88	A
DipInfo-UniTo	6.88	A
RDFpyrealb	6.86	A
SaarLST	6.83	A
OSU-CompLing	6.68	B
WebNLG-Human	6.67	B
DCU-NLG-Small	6.42	C

(e) Grammaticality human-en-D2T-1-FA

	Mean	Group
DCU-ADAPT-modPB	6.39	A
DCU-NLG-PBN	6.11	B
SaarLST	6.07	B
DipInfo-UniTo	6.01	B
OSU-CompLing	5.59	C
DCU-NLG-Small	5.51	C
WebNLG-Human	5.43	C
RDFpyrealb	4.53	D

(f) Grammaticality llm-en-D2T-1-FA

	Mean	Group
DCU-ADAPT-modPB	6.99	A
DCU-NLG-PBN	6.99	A
DipInfo-UniTo	6.96	A
SaarLST	6.95	AB
DCU-NLG-Small	6.87	BC
OSU-CompLing	6.82	CD
WebNLG-Human	6.77	D
RDFpyrealb	6.13	E

(g) Fluency human-en-D2T-1-FA

	Mean	Group
DCU-ADAPT-modPB	6.29	A
DCU-NLG-PBN	6.04	B
SaarLST	5.98	B
DipInfo-UniTo	5.89	B
OSU-CompLing	5.61	C
DCU-NLG-Small	5.50	C
WebNLG-Human	5.41	C
RDFpyrealb	4.69	D

(h) Fluency llm-en-D2T-1-FA

	Mean	Group
DCU-ADAPT-modPB	6.99	A
DCU-NLG-PBN	6.96	A
SaarLST	6.94	AB
DipInfo-UniTo	6.92	AB
OSU-CompLing	6.86	BC
DCU-NLG-Small	6.80	CD
WebNLG-Human	6.75	D
RDFpyrealb	6.10	E

Figure 6: System rankings for English in-domain factual data (left: human ratings, right: llm ratings)

publicly available for a few years and have been "ingested" by the different language models.

In-domain factual: Semantic accuracy criteria. In the human evaluation results, for `No-Omissions` SaarLST is in the first group with RDFpyrealb, while for `No-Additions`, most LLMs are in the first group, closely followed by RDFpyrealb. In both cases, OSU-CompLing, DCU-NLG-Small and the human-written texts stand at the bottom in the same group. In the LLM-as-judge evaluation, human-written texts are ranked in the first group for `No-Omissions`, and the middle one for `No-Additions`. The difference between human and LLM evaluation is rather important when it comes to evaluating the semantic accuracy of human-written texts.

In-domain factual: Intrinsic quality criteria. In the human evaluation, for both criteria, DCU-ADAPT-modPB ranks alone in the first group, followed by DCU-NLG-PBN, SaarLST and DipInfo-UniTo in the second group, OSU-CompLing, DCU-NLG-Small and human-written texts in the third group, and RDFpyrealb in the fourth group. Results are less clear cut in the LLM-as-judge evaluation with the same absolute score rankings but with some overlaps between the groups.

Comments on in-domain factual results. In previous similar multi-system evaluations of

(a) No-omissions human-en-*-*

	Mean	Group
SaarLST	5.89	A
DipInfo-UniTo	5.58	B
RDFpyrealb	5.57	B
DCU-NLG-PBN	5.43	C
OSU-CompLing	4.73	D
DCU-NLG-Small	4.47	E

(b) No-omissions llm-en-*-*

	Mean	Group
SaarLST	6.85	A
RDFpyrealb	6.78	A
DipInfo-UniTo	6.57	B
DCU-NLG-PBN	6.56	B
OSU-CompLing	6.19	C
DCU-NLG-Small	5.46	D

(c) No-additions human-en-*-*

	Mean	Group
SaarLST	5.66	A
DipInfo-UniTo	5.60	A
DCU-NLG-PBN	5.31	B
RDFpyrealb	5.24	B
OSU-CompLing	4.56	C
DCU-NLG-Small	4.31	D

(d) No-additions llm-en-*-*

	Mean	Group
DipInfo-UniTo	6.83	A
SaarLST	6.80	AB
DCU-NLG-PBN	6.74	B
RDFpyrealb	6.73	B
OSU-CompLing	6.60	C
DCU-NLG-Small	5.92	D

(e) Grammaticality human-en-*-*

	Mean	Group
SaarLST	6.04	A
DipInfo-UniTo	5.85	B
DCU-NLG-PBN	5.79	B
OSU-CompLing	5.15	C
DCU-NLG-Small	5.09	C
RDFpyrealb	4.44	D

(f) Grammaticality llm-en-*-*

	Mean	Group
SaarLST	6.97	A
DCU-NLG-PBN	6.95	AB
DipInfo-UniTo	6.92	B
DCU-NLG-Small	6.75	C
OSU-CompLing	6.75	C
RDFpyrealb	5.88	D

(g) Fluency human-en-*-*

	Mean	Group
SaarLST	5.98	A
DipInfo-UniTo	5.78	B
DCU-NLG-PBN	5.75	B
OSU-CompLing	5.23	C
DCU-NLG-Small	5.07	D
RDFpyrealb	4.60	E

(h) Fluency llm-en-*-*

	Mean	Group
SaarLST	6.95	A
DCU-NLG-PBN	6.93	A
DipInfo-UniTo	6.87	B
OSU-CompLing	6.82	C
DCU-NLG-Small	6.61	D
RDFpyrealb	5.85	E

Figure 7: System rankings for English overall (left: human ratings, right: llm ratings)

data-to-text generation on factual in-domain data, i.e WebNLG'17 (Gardent et al., 2017) WebNLG'20 (Castro Ferreira et al., 2020) and WebNLG'23 (Cripwell et al., 2023), the human-written texts were in the first or occasionally second group. In our evaluation, human-written texts rank in the third or fourth group depending on the criterion. Given that LLMs are now able to produce very natural texts and that, to ensure semantic accuracy, original WebNLG texts were created under a set of constraints possibly limiting the naturalness of the texts, seeing human-written texts getting behind LLMs on `Grammaticality` and `Fluency` can be expected. What could be considered more surprising is the fact that in terms of semantic accuracy, the 2020 human-written texts are now ranked below RDFpyrealb, the rule-based system whose outputs were also submitted in 2020. Although it is possible that RDFpyrealb was improved beyond human quality in terms of semantic accuracy, this could also be an indicator that ranking-based evaluation results such as the one presented here are eventually relative to the current state of the art, as noted recently in the speech synthesis domain (Le Maguer et al., 2024).

4.1.4 Results for English across all datasets

Figure 7 shows the overall rankings on the English data. The tables summarize what has been described in the previous sections: SaarLST consistently ranks first for all criteria, followed by DipInfo-UniTo and DCU-NLG-PBN (DipInfo-UniTo being better on semantic accuracy criteria), then OSU-CompLing and DCU-NLG-Small (OSU-CompLing being better on `Grammaticality`). RDFpyrealb ranks in the second cluster for semantic accuracy criteria, and last for `Grammaticality` and `Fluency`. DCU-NLG-PBN manages to reach a level close to that of larger or multiple LLMs with one single 7B-instruct model.

4.1.5 Takeaways from English results

Having six different test sets, a variety of system implementations, four different quality criteria and several evaluation methods allow us to get these interesting insights on the results.

There is no degradation of scores on out-of-domain data except for rule-based systems. One possible explanation is that LLMs have all been exposed to Wikipedia texts, from which the Wikidata triples we collected for the out-of-domain datasets generally come from. But the fully rule-based system (RDFpyrealb) is the only one that does not degrade on counterfactual and fictional data. The overall score degradation of all systems is rather moderate on counterfactual and fictional data.

LLMs give rankings that look consistent with human rankings, with a couple of notable exceptions. First, LLMs tend to rank the fully rule-based system (RDFpyrealb) higher than humans do on semantic accuracy criteria. This could be due to the fact that humans are more impacted by the naturalness of the produced sentences when evaluating semantic accuracy (RDFpyrealb systematically ranks at the bottom for intrinsic quality criteria). Note that LLMs also rank higher human-written texts, which are also of lower quality in terms of `Grammaticality` and `Fluency` according to both human and LLM-as-judge evaluations. A second and more curious result, DCU-NLG-PBN is also generally ranked higher by LLMs than by humans on the intrinsic quality criteria. One plausible explanation for this anomaly could be that the output from DCU-NLG-PBN is structured in way that has greater alignment with the evaluation criteria and outputs that the model evaluators have already seen. Results from LLM evaluators can vary between across datasets and properties being judged (Bavaresco et al., 2025). See Section 4.2 for a detailed analysis of correlations.

Both humans and LLMs assign higher scores for intrinsic quality criteria (`Grammaticality` and `Fluency`) than for semantic accuracy criteria (`No-Omissions` and `No-Additions`). This could be an indication that semantic accuracy is more difficult to handle for systems across the board; it is also possible that semantic accuracy is more difficult to assess, since aligning precisely the semantics of texts and input tables is a challenge that is still to be solved.

LLMs assign much higher scores and produce more ties overall than humans to all outputs. By looking at the raw evaluation results (not show here), it is striking that LLMs very often assign maximal scores of 7, which is not the case with human evaluators. The absolute text quality ratings provided my LLMs need to be taken cautiously. We also counted the ties across all English results tables (Figures 1 to 7): reading the tables from top to bottom, we counted the number of times a system is placed in the same group as another system, which happens 73 times in human tables, and 89 times in the LLM-as-judge tables.

General comments on the systems. Overall, systems using more resources usually rank higher, and fine-tuned Mistral-7B seems to perform better than fine-tuned LLama-7B on the task. A comparison between RDFpyrealb and DCU-NLG-Small is also interesting. Both use handwritten grammars as main generation component, but DCU-NLG-Small adds a paraphrasing component to improve the intrinsic quality of the text, which has traditionally been challenging for rule-based systems. DCU-NLG-Small gets better results than RDFpyrealb in terms of those criteria, occasionally ranking in the same group as an LLM-based submission, but this comes at the expense of semantic accuracy, for which DCU-NLG-Small consistently ranks at the bottom, while RDFpyrealb is often on par with or close to the best systems.

System-level correlations on all results presented in this section are provided in Section 4.3, while instance-level correlations on individual datasets and overall are presented in Section 4.2.

4.1.6 Takeways from Spanish results

The results of the human and LLM-as-judge evaluations are shown in Figures 8 to 13. For Spanish data, there are only three systems and the picture is quite clearer than for English, so we do not break the analysis down into subsections.

It is preferable to directly fine-tune a Spanish model than to fine-tune an English model and machine translate its output. OSU-CompLing, which is a fine-tuned Spanish LLama2-7B model, is systematically in the first group according to both human and LLM-as-judge scores (with only one exception, LLM's `No-Addition` table for in-domain data, in Figure 8). The DCU-NLG-PBN scores are quite close to OSU-CompLing's, and the rankings place it most of the times in the first group as well, and sometimes in the second group. Given that (i) DCU-NLG-PBN used a heavier pipeline, which consists of a fine-tuned Mistral-7B model that gen-

(a) No-omissions human-es-D2T-1-*

	Mean	Group
OSU-CompLing	6.09	A
DCU-NLG-PBN	5.88	B
DCU-NLG-Small	4.93	C

(b) No-omissions llm-es-D2T-1-*

	Mean	Group
OSU-CompLing	6.74	A
DCU-NLG-PBN	6.50	B
DCU-NLG-Small	5.75	C

(c) No-additions human-es-D2T-1-*

	Mean	Group
OSU-CompLing	5.79	A
DCU-NLG-PBN	5.73	A
DCU-NLG-Small	4.75	B

(d) No-additions llm-es-D2T-1-*

	Mean	Group
DCU-NLG-PBN	6.83	A
OSU-CompLing	6.74	B
DCU-NLG-Small	6.31	C

(e) Grammaticality human-es-D2T-1-*

	Mean	Group
OSU-CompLing	6.66	A
DCU-NLG-PBN	6.56	B
DCU-NLG-Small	6.01	C

(f) Grammaticality llm-es-D2T-1-*

	Mean	Group
OSU-CompLing	6.97	A
DCU-NLG-PBN	6.96	A
DCU-NLG-Small	6.83	B

(g) Fluency human-es-D2T-1-*

	Mean	Group
OSU-CompLing	6.62	A
DCU-NLG-PBN	6.51	B
DCU-NLG-Small	5.95	C

(h) Fluency llm-es-D2T-1-*

	Mean	Group
OSU-CompLing	6.97	A
DCU-NLG-PBN	6.95	A
DCU-NLG-Small	6.75	B

Figure 8: System rankings for Spanish in-domain data (left: human ratings, right: llm ratings)

(a) No-omissions human-es-D2T-2-*

	Mean	Group
OSU-CompLing	6.04	A
DCU-NLG-PBN	5.86	B
DCU-NLG-Small	4.49	C

(b) No-omissions llm-es-D2T-2-*

	Mean	Group
OSU-CompLing	6.74	A
DCU-NLG-PBN	6.58	B
DCU-NLG-Small	5.26	C

(c) No-additions human-es-D2T-2-*

	Mean	Group
OSU-CompLing	5.62	A
DCU-NLG-PBN	5.56	A
DCU-NLG-Small	4.12	B

(d) No-additions llm-es-D2T-2-*

	Mean	Group
OSU-CompLing	6.73	A
DCU-NLG-PBN	6.71	A
DCU-NLG-Small	5.71	B

(e) Grammaticality human-es-D2T-2-*

	Mean	Group
DCU-NLG-PBN	6.58	A
OSU-CompLing	6.57	A
DCU-NLG-Small	5.54	B

(f) Grammaticality llm-es-D2T-2-*

	Mean	Group
OSU-CompLing	6.98	A
DCU-NLG-PBN	6.97	A
DCU-NLG-Small	6.77	B

(g) Fluency human-es-D2T-2-*

	Mean	Group
OSU-CompLing	6.55	A
DCU-NLG-PBN	6.54	A
DCU-NLG-Small	5.50	B

(h) Fluency llm-es-D2T-2-*

	Mean	Group
DCU-NLG-PBN	6.97	A
OSU-CompLing	6.97	A
DCU-NLG-Small	6.61	B

Figure 9: System rankings for Spanish out-of-domain data (left: human ratings, right: llm ratings)

(a) No-omissions human-es-*-FA

	Mean	Group
OSU-CompLing	6.10	A
DCU-NLG-PBN	5.95	A
DCU-NLG-Small	4.84	B

(b) No-omissions llm-es-*-FA

	Mean	Group
OSU-CompLing	6.79	A
DCU-NLG-PBN	6.58	B
DCU-NLG-Small	5.61	C

(c) No-additions human-es-*-FA

	Mean	Group
DCU-NLG-PBN	5.88	A
OSU-CompLing	5.78	A
DCU-NLG-Small	4.68	B

(d) No-additions llm-es-*-FA

	Mean	Group
DCU-NLG-PBN	6.81	A
OSU-CompLing	6.76	A
DCU-NLG-Small	6.07	B

(e) Grammaticality human-es-*-FA

	Mean	Group
OSU-CompLing	6.70	A
DCU-NLG-PBN	6.69	A
DCU-NLG-Small	5.85	B

(f) Grammaticality llm-es-*-FA

	Mean	Group
DCU-NLG-PBN	6.97	A
OSU-CompLing	6.97	A
DCU-NLG-Small	6.76	B

(g) Fluency human-es-*-FA

	Mean	Group
OSU-CompLing	6.66	A
DCU-NLG-PBN	6.66	A
DCU-NLG-Small	5.80	B

(h) Fluency llm-es-*-FA

	Mean	Group
OSU-CompLing	6.97	A
DCU-NLG-PBN	6.97	A
DCU-NLG-Small	6.66	B

Figure 10: System rankings for Spanish factual data (left: human ratings, right: llm ratings)

(a) No-omissions human-es-*-CFA

	Mean	Group
OSU-CompLing	5.99	A
DCU-NLG-PBN	5.79	B
DCU-NLG-Small	4.68	C

(b) No-omissions llm-es-*-CFA

	Mean	Group
OSU-CompLing	6.64	A
DCU-NLG-PBN	6.40	B
DCU-NLG-Small	5.30	C

(c) No-additions human-es-*-CFA

	Mean	Group
OSU-CompLing	5.50	A
DCU-NLG-PBN	5.40	A
DCU-NLG-Small	4.22	B

(d) No-additions llm-es-*-CFA

	Mean	Group
OSU-CompLing	6.67	A
DCU-NLG-PBN	6.64	A
DCU-NLG-Small	5.86	B

(e) Grammaticality human-es-*-CFA

	Mean	Group
OSU-CompLing	6.52	A
DCU-NLG-PBN	6.49	A
DCU-NLG-Small	5.64	B

(f) Grammaticality llm-es-*-CFA

	Mean	Group
OSU-CompLing	6.97	A
DCU-NLG-PBN	6.94	A
DCU-NLG-Small	6.74	B

(g) Fluency human-es-*-CFA

	Mean	Group
OSU-CompLing	6.47	A
DCU-NLG-PBN	6.42	A
DCU-NLG-Small	5.58	B

(h) Fluency llm-es-*-CFA

	Mean	Group
OSU-CompLing	6.95	A
DCU-NLG-PBN	6.95	A
DCU-NLG-Small	6.61	B

Figure 11: System rankings for Spanish counterfactual data (left: human ratings, right: llm ratings)

(a) No-omissions human-es-*-FI

	Mean	Group
OSU-CompLing	6.11	A
DCU-NLG-PBN	5.87	B
DCU-NLG-Small	4.61	C

(b) No-omissions llm-es-*-FI

	Mean	Group
OSU-CompLing	6.79	A
DCU-NLG-PBN	6.65	B
DCU-NLG-Small	5.60	C

(c) No-additions human-es-*-FI

	Mean	Group
OSU-CompLing	5.84	A
DCU-NLG-PBN	5.66	A
DCU-NLG-Small	4.41	B

(d) No-additions llm-es-*-FI

	Mean	Group
DCU-NLG-PBN	6.85	A
OSU-CompLing	6.79	A
DCU-NLG-Small	6.10	B

(e) Grammaticality human-es-*-FI

	Mean	Group
OSU-CompLing	6.63	A
DCU-NLG-PBN	6.53	A
DCU-NLG-Small	5.82	B

(f) Grammaticality llm-es-*-FI

	Mean	Group
OSU-CompLing	6.98	A
DCU-NLG-PBN	6.97	A
DCU-NLG-Small	6.89	B

(g) Fluency human-es-*-FI

	Mean	Group
OSU-CompLing	6.61	A
DCU-NLG-PBN	6.50	A
DCU-NLG-Small	5.79	B

(h) Fluency llm-es-*-FI

	Mean	Group
OSU-CompLing	6.98	A
DCU-NLG-PBN	6.98	A
DCU-NLG-Small	6.78	B

Figure 12: System rankings for Spanish fictional data (left: human ratings, right: llm ratings)

(a) No-omissions human-es-*-*

	Mean	Group
OSU-CompLing	6.07	A
DCU-NLG-PBN	5.87	B
DCU-NLG-Small	4.71	C

(b) No-omissions llm-es-*-*

	Mean	Group
OSU-CompLing	6.74	A
DCU-NLG-PBN	6.54	B
DCU-NLG-Small	5.50	C

(c) No-additions human-es-*-*

	Mean	Group
OSU-CompLing	5.71	A
DCU-NLG-PBN	5.65	A
DCU-NLG-Small	4.44	B

(d) No-additions llm-es-*-*

	Mean	Group
DCU-NLG-PBN	6.77	A
OSU-CompLing	6.74	A
DCU-NLG-Small	6.01	B

(e) Grammaticality human-es-*-*

	Mean	Group
OSU-CompLing	6.61	A
DCU-NLG-PBN	6.57	A
DCU-NLG-Small	5.77	B

(f) Grammaticality llm-es-*-*

	Mean	Group
OSU-CompLing	6.97	A
DCU-NLG-PBN	6.96	A
DCU-NLG-Small	6.80	B

(g) Fluency human-es-*-*

	Mean	Group
OSU-CompLing	6.58	A
DCU-NLG-PBN	6.53	A
DCU-NLG-Small	5.72	B

(h) Fluency llm-es-*-*

	Mean	Group
OSU-CompLing	6.97	A
DCU-NLG-PBN	6.96	A
DCU-NLG-Small	6.68	B

Figure 13: System rankings for Spanish overall (left: human ratings, right: llm ratings)

erates English outputs and the Google Translate API[10] to produce Spanish outputs, and (ii) DCU-NLG-PBN consistently ranked higher than OSU-CompLing in English with the same models, it seems preferable to fine-tune language-specific models rather than to use machine translation. DCU-NLG-Small, which also uses machine translation (NLLB (Team et al., 2022)) on the English outputs, is in the last group for all criteria and on all datasets (second group when OSU-CompLing and DCU-NLG-PBN are tied, third group when they are not).

LLMs are robust on out-of-domain and fictional data, but possibly not as much on counterfactual data. OSU-CompLing and DCU-NLG-PBN are generally robust on out-of domain data (with maybe a small score drop for the `No-Additions` criterion), while DCU-NLG-Small suffers a more important score decrease for all criteria. On counterfactual data, all systems see their respective scores decrease for all four criteria, with only the LLM-as-judge ratings of `Grammaticality` and `Fluency` being at the same level. As it was the case for English, the systems look more robust on the fictional dataset, but here too only the system with a rule-based component (DCU-NLG-Small) does not see its scores drop for `Grammaticality` and `Fluency`. For DCU-NLG-Small, although humans give it lower scores on the counterfactual data for the semantic accuracy criteria compared to the factual dataset, LLMs assign very similar scores on both datasets.

In the overall results in Figure 13, both humans and LLMs rank jointly OSU-CompLing and DCU-NLG-PBN in the first group for all criteria but `No-Omissions`, for which DCU-NLG-PBN is ranked second; DCU-NLG-Small is always alone in the last group. The lower scores of DCU-NLG-PBN for `No-Omissions` could be due to a lack of robustness on the counterfactual and fictional subsets (see Figures 11 and 12).

LLMs and humans score different but rank the same. As it was the case for English, LLMs tend to score all systems higher than humans, but the system rankings are largely aligned with the human system rankings. Sections 4.2 and 4.3 provide more in-depth analysis of the correlations between the different evaluation methods.

[10] https://cloud.google.com/translate

4.2 Instance-level correlations between human and LLM-as-judge evaluations

English and Spanish bird's eye view system results. For the data-to-text system results, there are several general patterns that are apparent across the different quality criteria in both the English and Spanish results. Firstly, there is a general divergence between the average human and LLM scores across all of the evaluation criteria. Across the different systems, the average LLM score most of the time is higher than the equivalent average human score as observed in section 4.1.5, with the LLMs giving higher ratings. For the English results the divergence is more acute for some systems than others e.g. RDFpyrealb, OSU-CompLing, and DCU-NLG-Small. However, this is not too surprising given that these systems find themselves at the bottom of the various system rankings for either some or most of the different quality criteria across the different datasets. For Spanish only the DCU-NLG-Small system has an acute divergence between the average human and LLM scores.

We plotted the LLM scores against the human scores for each criterion (see Appendix E). These plots show clearly that whilst the LLMs consistently rate higher than the human annotators, they seem agree much more with one another in terms of the intrinsic quality criteria (systems are grouped more compactly on the horizontal axis than for the semantic accuracy criteria). The English results (Figures 22 and 23) differ from the Spanish results (Figures 26 and 27) in that there is a greater uniformity between the LLM scores over the different systems (except DCU-NLG-Small) compared to the English ratings. It is possible the reason for the greater uniformity of results for the Spanish system outputs could be the small amount of systems evaluated (three), but it could also be due to a higher quality of the annotators employed (e.g. bi-lingual skills), or it could be that the Spanish annotators have used LLMs in assessing the outputs.

English human-LLM correlations. We analysed the consistency of ratings between LLMs and humans across the different evaluation criteria. Figure 14 shows a comprehensive correlation matrix of aggregated scores across all systems, models, and evaluation dimensions for English (see Section 3.4 for details about the computation of the correlations). At first sight, two darker square are clearly visible, on the one hand the correlation scores between all evaluators on

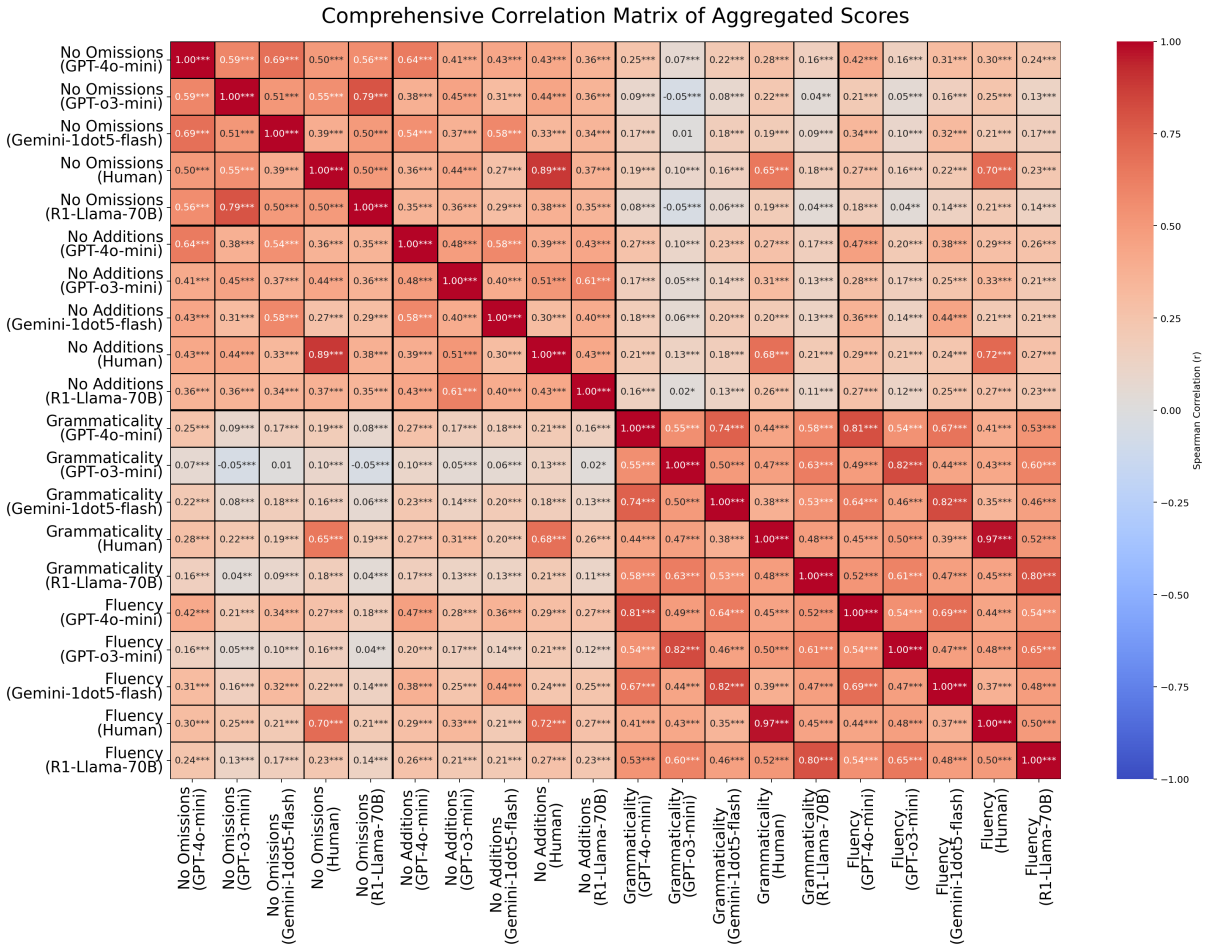


Figure 14: Instance-level correlations over all English scores (all datasets, all systems).

No-Omissions and No-Additions, and on the other hand the correlations between all evaluators on Grammaticality and Fluency. This suggests that (i) the semantic accuracy criteria group and the intrinsic quality criteria group are complementary aspects of quality, and (ii) evaluators may have difficulties in distinguishing between criteria within one group, or that the systems that are good according to one criterion of one group are also good at the other criterion within the same group.[11]

With intrinsic quality criteria like Fluency and Grammaticality, scores across GPT-4, GPT-3.5, LLaMA, and human evaluations are indeed positively correlated ($0.4 > \rho > 0.5$ in most cases), meaning that while AI evaluators generally agree with humans, they are not perfect substitutes. The results suggest promising alignment, while underscoring that human assessments still identify nuances often overlooked by models.

For the No-Additions and No-Omissions dimensions, LLMs and humans are also positively correlated ($0.4 > \rho > 0.5$ in most cases). Their correlations with Fluency and Grammaticality are weaker ($\rho < 0.4$) and sometimes even negative. These dimensions capture complementary aspects of quality that are not fully reflected in Fluency or Grammaticality scores. Note however that human Grammaticality and human Fluency have high correlations with human No-Omissions and human No-Additions (darker cells in the lower left and upper right squares), which is consistent with our above observation about the scores, and could indicate that the human assessment of one group of criteria (e.g. No-Omissions or No-Additions) is impacted by the output quality in terms of the other group of criteria (e.g. Fluency or Grammaticality).

Finally, as shown in the matrices computed sepa-

[11] The latter would be supported by the results seen in Section 4.1, in which we saw a system like DCU-NLG small (or DCU-ADAPT-modPB) which has different scores between No-Omissions and No-Additions, but similar scores for Grammaticality and Fluency: in Figure 14 the colour of the Grammaticality/Fluency is slightly darker that the colour of the No-Omissions/No-Additions square.

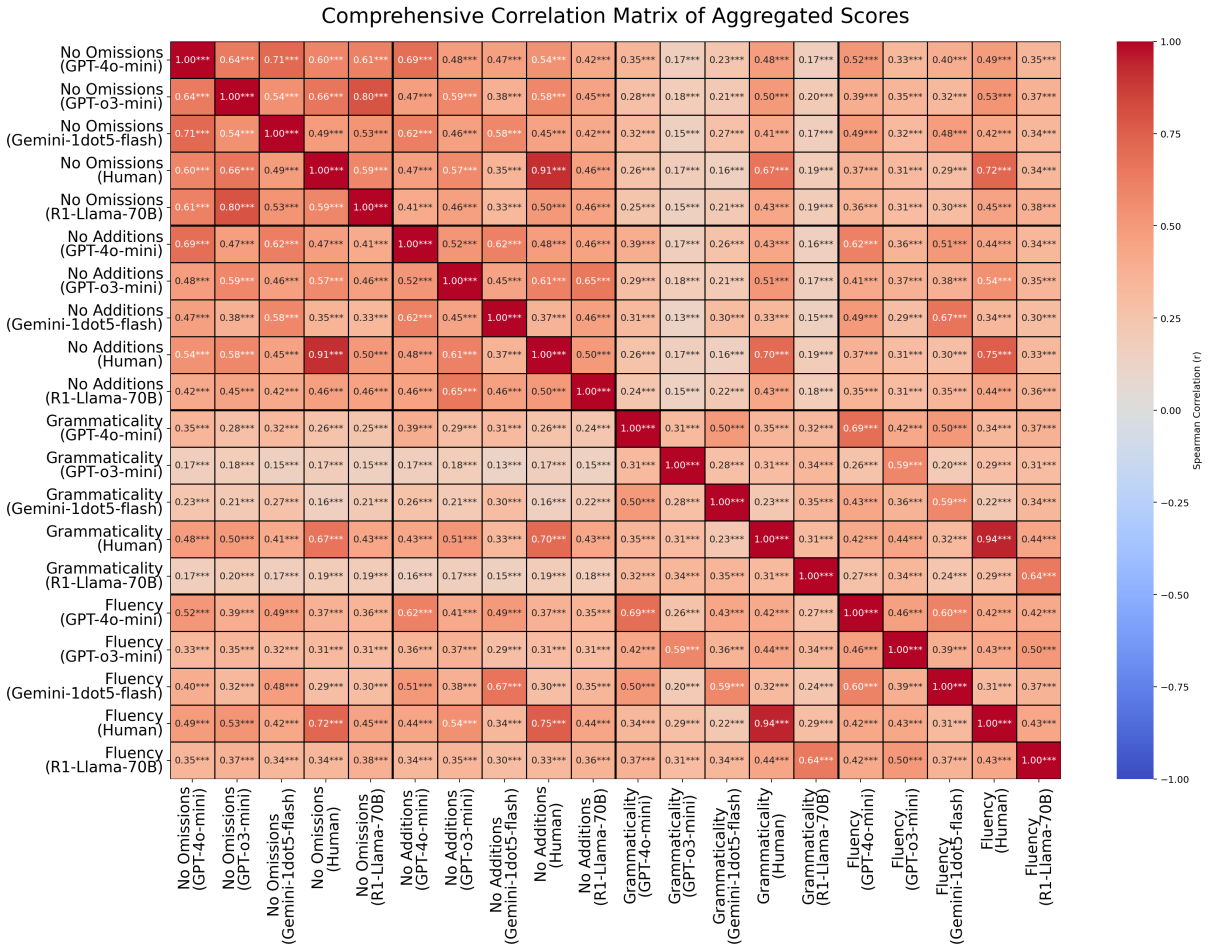


Figure 15: Instance-level correlations over all Spanish scores (all datasets, all systems).

rately for the size different datasets (see Figure 20 in Appendix D), which are visually very similar to Figure 14, there does not seem to be major differences in terms of instance-level correlations across the datasets.

Spanish human-LLM correlations. In comparison to the English correlation results, the Spanish matrix looks at first sight more homogenous (Figure 15), with the squares being less visible and fewer correlations below 0.2. `No-Omissions` and `No-Additions` correlations between LLMs and humans are higher than in English ($0.5 > \rho > 0.7$ in most cases), while `Fluency` and `Grammaticality` show weaker correlations between LLMs and humans overall ($0.3 > \rho > 0.4$ in most cases). Unlike in English, all `No-Omissions` and `No-Additions` scores have rather high correlations with human `Fluency` and `Grammaticality` scores ($0.4 > \rho > 0.7$ in most cases). This is difficult to interpret given that only three systems were evaluated in Spanish.

Note that as in English, correlations between human `No-Omissions`/`No-Additions` and human `Fluency`/`Grammaticality` are strong, at around 0.7, and the correlation between `No-Omissions` and `No-Additions` is even stronger at 0.89, in the same fashion as the correlation between `Fluency` and `Grammaticality` at 0.97. Also as it was the case for English, each dataset-specific matrix (see Figure 21 in Appendix D) is very similar to the overall matrix in Figure 15.

4.3 System-level correlations between human, metric and LLM-as-judge evaluations

We computed Spearman's rank correlations on all system rankings[12] according to all metrics, LLMs and human ratings. Figures 16 to 18 show the results. Cells are annotated with ρ_{ab} and significance stars for $q_{ab} < 0.05, 0.01, 0.001$ (shown as $*, **, ***$). The heatmap visualizes ρ_{ab} on a fixed $[-1, 1]$ diverging scale. See Section 3.4 for details.

[12]For better comparability across matrices, we only used the same 6 systems that submitted outputs for all datasets; in other words, we did not include DCU-ADAPT-modPB for computing the rank correlations.

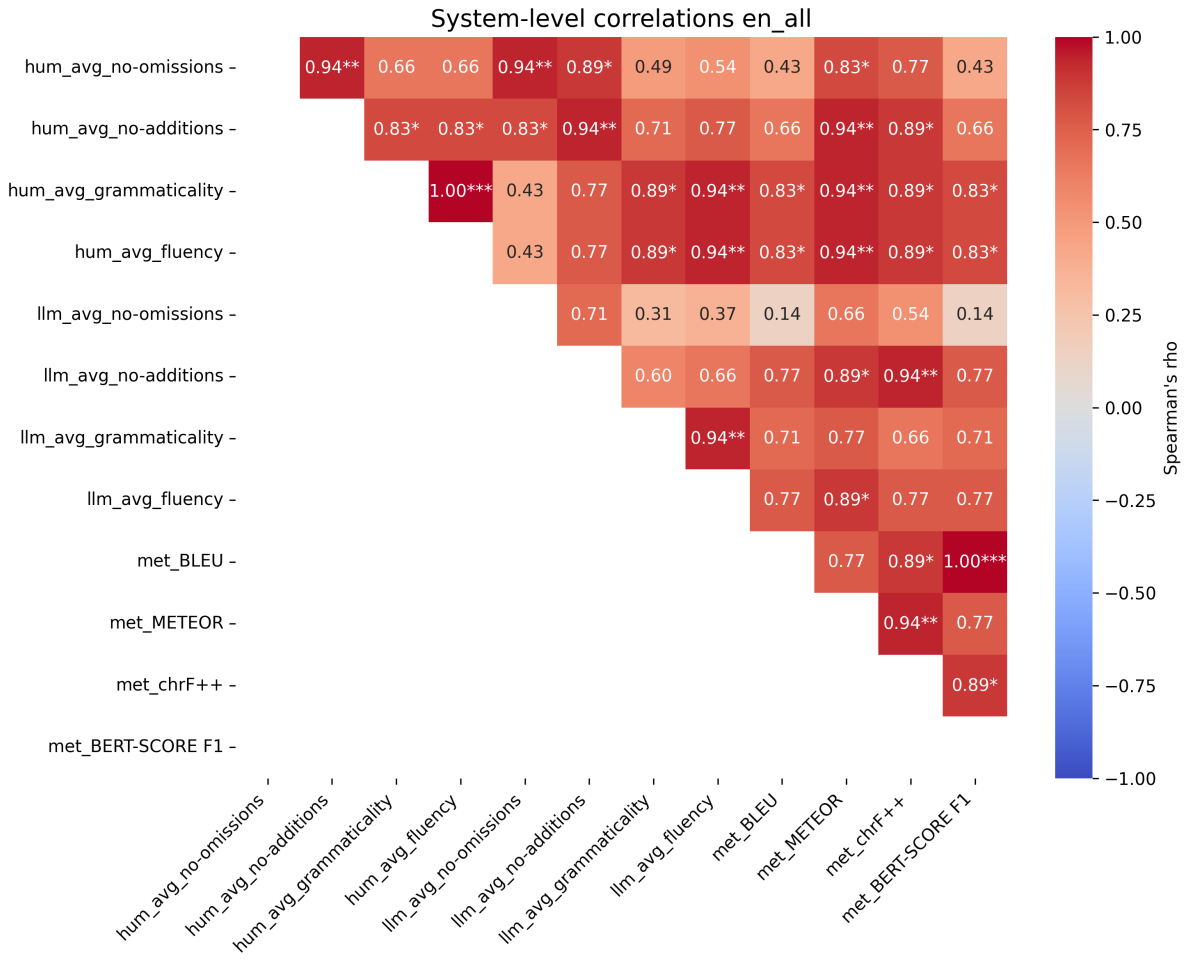


Figure 16: English overall data-to-text Spearman's system-level correlations

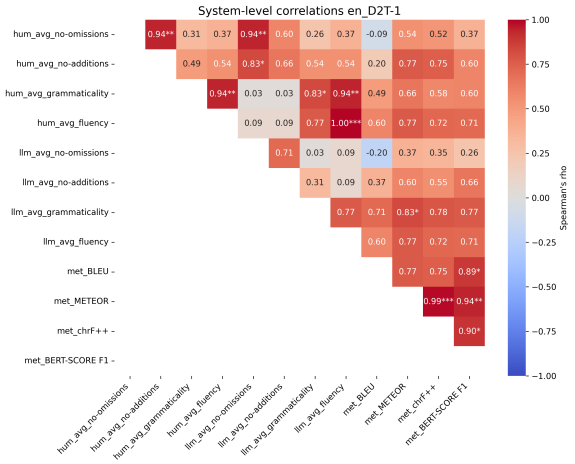
Comparing LLMs and humans. When looking at the overall Spearman correlations (Figure 16) we can that the LLMs evaluations correlate statistically positively with their human equivalent quality criterion for all of four criteria: 0.89 for `Grammaticality`, and 0.94 for `No-Omissions`, `No-Additions` and `Fluency` (all at $q_{ab} < 0.01$); this holds across nearly all of the datasets.

Comparing different human quality assessments. In the analysis across all of the English datasets (Figure 16) we see that humans, unlike LLMs, show a statistically positive correlation between their `No-Omissions` and `No-Additions` system rankings (0.94, $q_{ab} < 0.01$); humans also exhibit perfect correlations between their `Grammaticality` and `Fluency` rankings.

Unlike LLMs, humans also show statistically positive $q_{ab} < 0.05$ correlations for `Fluency` and `Grammaticality` with the human `No-Additions` criterion. Whilst there is a positive correlation for these intrinsic quality criteria with the human `No-Omissions`, this is not seen as statistically significant. However, on each of the dataset-specific analyses (Figures 17a to 18c) there are variances with some datasets not showing any statistically significant correlations (Figures 17a, 18b and 18c), partial statistical correlations (Figures 18a, to complete statistical correlation of all evaluation dimensions (Figure 17b).

Comparing human and LLM against automatic metrics. Finally, we also explored the correlation between the human and LLM evaluation scores against those from established automatic metrics. In particular, we use for our comparison a combination of text overlap metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and chrF++ (Popović, 2017). Additionally, we included BERTScore (Zhang et al., 2019) as a semantic similarity metric. With the exception of chrF++, these are some of the popular automatic metrics within natural language generation (Schmidtova et al., 2024).

(a) English D2T-1 Spearman's system-level correlations



(b) English D2T-2 Spearman's system-level correlations

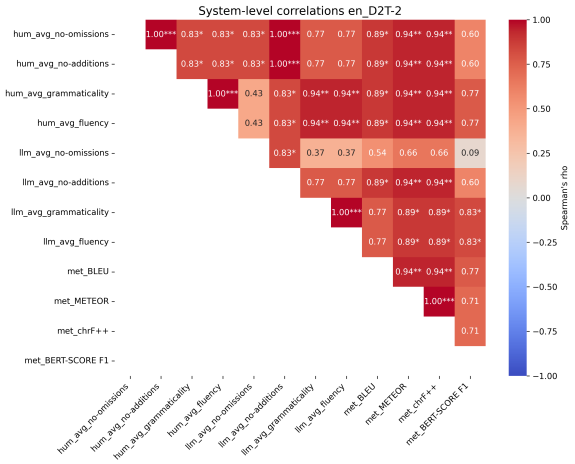


Figure 17: System-level correlations: In-domain (D2T-1) and out-of-domain (D2T-2) data)

We find that indeed LLMs are more statistically significantly correlated to human judgements than automatic metrics. However, both METEOR and chrF++ have statistically significant system-level ranking correlations to human judgements of `Fluency` and `Grammaticality`. What is surprising is that LLM and automatic metrics are much lower in correlation than either with humans. We see this pattern across all datasets.

Looking at the per-dataset analyses variations in the degree of correlation between the automatic metrics and both the human and LLM average metric results. For some datasets there are no statistical correlations (Figures 17a) or partial human metric correlations (Figures 18a, 18b, and 18c). Only the English D2T-2 dataset (Figure 17b) shows complete positive statistical correlation for human scores and most of the LLM scores against the overlap metrics. Interestingly enough, for the same dataset the semantic based BERTScore does not show any statistically positive correlations for all of the human and most of the LLM scores.

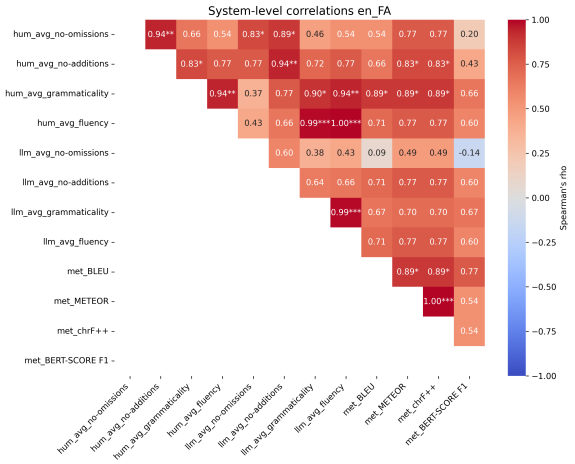
Traditionally, lexical automatic metrics were only used at the system-level (Papineni et al., 2002), but these have been used at an instance-level (Liu et al., 2016). BLEU has been shown not to reliably predict human judgments, but is possibly useful at a system-level. Note that BLEU has clearly higher correlations with human `Grammaticality` and human `Fluency` than with the semantic accuracy criteria, which is expected since it is an n-gram-based metric, which is by definition more surface-oriented. However, more surprisingly, we do not observe a positive correlation between the embedding-based (thus content-oriented) BERTScore and these semantic accuracy criteria, while BERTScore does correlate positively with both intrinsic quality criteria. In our experiment BLEU and BERTScore even have a system-level correlation of 1.0 on the English outputs.

The curious case of the D2T-2-CFA scores. In (Mille et al., 2024b), we pointed out the unexpectedly high metrics scores (BLEU, METEOR, chrF++, BERTScore) obtained by almost all systems on the out-of-domain counterfactual data (D2T-2-CFA). When looking at Table 4 in Appendix B, we observe that almost all human scores (and most LLM scores) for all systems are lower on D2T-2-CFA than on D2T-2-FA, which may indicate that there is a quality problem with the D2T-2-FA and/or D2T-2-CFA reference texts we collected for computing the metrics scores. Unlike the human scores, the LLM scores for `Grammaticality` and `Fluency` tend to be at the same level as the corresponding D2T-2-FA scores; it could be the case that either or both the writing and the evaluation of D2T-2-CFA texts by humans are somewhat challenging. More research is needed on the topic to find out what is exactly happening.

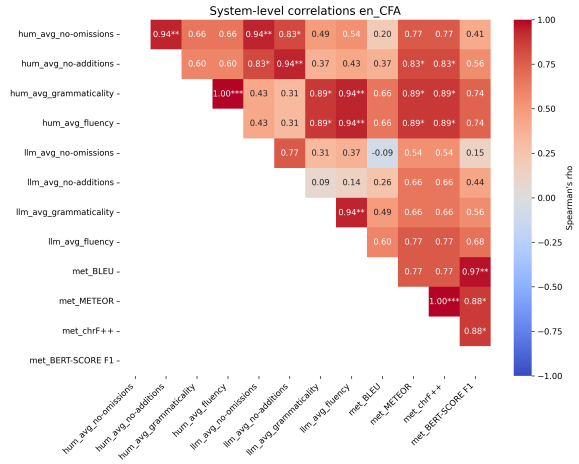
4.4 Comments on human annotators

By using annotator training and filtering annotators based on agreement levels, we were able to find annotators with high levels of agreement. On the English data, we found Krippendorff's alpha internal based agreement levels of 0.64, 0.67, 0.47, 0.43 for `No-Omissions`, `No-Additions`, `Grammaticality`, and `Fluency`

(a) English D2T-FA Spearman's system-level correlations



(b) English D2T-CFA Spearman's system-level correlations



(c) English D2T-FI Spearman's system-level correlations

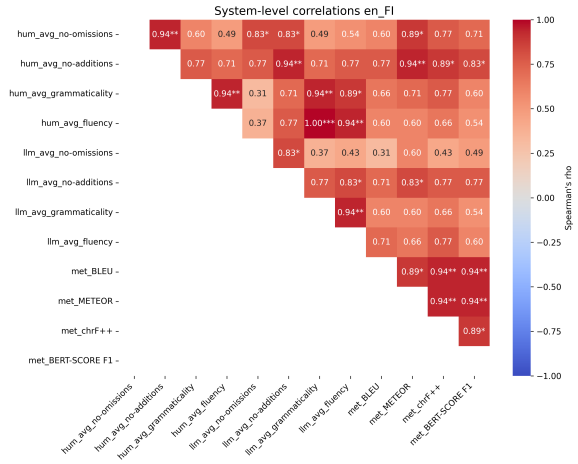


Figure 18: System-level correlations: Factual (FA), Counterfactual (CFA) and Fictional (FI) data)

respectively. Although we had a smaller number of annotators for Spanish, the level of agreement was higher at 0.84, 0.85, 0.72, 0.72 for `No-Omissions`, `No-Additions`, `Grammaticality`, and `Fluency` respectively. There were no significant differences in annotator agreement levels across data subsets. However, as shown in (Zhang et al., 2023) and other works, high agreement does not necessarily mean that the annotations agree correctly.

Bias Considerations. This analysis is rank-based, which reduces sensitivity to scale differences across models. Evaluator separation ensures that judgments are not accidentally double-counted. However, because stacking was used, evaluators with more items contribute more weight, which may bias correlations toward heavily represented annotators. Pairwise deletion further implies that n varies by cell; if missingness is not random, this may distort estimates. Moreover, p-values are not corrected for multiple comparisons, and constant-score models can yield unstable or undefined correlations. Despite these caveats, the matrices provide an informative overview of system-level agreement patterns across criteria.

5 Conclusions

From the evaluation results in section 4 it seems that there is an abundantly clear pattern that can be seen for the English results. Those systems that use more resources in the form of larger or multiple models tend to outperform the smaller system implementations whether they be purely rule-based (RDFpyrealb), hybrid neural-symbolic (DCU-NLG-Small), or just use a smaller fine-tuned LLM model end-to-end (OSU-CompLing). It is worth noting the singular exception; the DCU-NLG-PBN system with its 7B fine-tuned model can match or exceed heavier implementations such as DCU-ADAPT-modPB which uses GPT-4.

When looking at the Spanish results the same

does not hold as true as for the English results. The OSU-CompLing system usually matches or exceeds the multiple model implementation of DCU-NLG-PBN system (Mistral 7B + Machine Translation). One factor for this difference could be due to the fact that it leverages a fine-tuned model for Spanish as opposed to generating in English first and then translating.

We also compared and contrasted the same evaluations conducted by humans and LLMs. We saw that both humans and LLMs are usually aligned in their rankings of the systems across the different quality criteria evaluated and also for both English and Spanish. This is encouraging and seems to indicate the possibility of using LLMs as a means to rank the output from different systems that would be similar to human preferences.

The very high mean scores assigned by LLMs, which often reach 6.9/7 and above, need to be put in perspective of the human evaluation results, which are typically lower and more conservative. This holds across both languages, different datasets, and the various evaluation criterion. There is certainty room from improvement in getting LLMs to score more like humans on Likert scales for semantic and intrinsic evaluation criterion.

Another observation that we have seen is that LLMs tend to produce more ties in its scoring than human evaluators. It remains to be investigated if it is because LLMs have more problems distinguishing between different outputs of similar quality, or because human scoring is too fine-grained that models are unable to replicate.

When looking at the correlations between the semantic and intrinsic quality criterion, we can see several interesting patterns. There is a strong positive correlation for humans between the semantic and intrinsic quality criteria. It is likely that for human evaluators the semantic accuracy scores are impacted by the intrinsic quality of the texts in both English and Spanish. This inter-dependency was not observed with LLM evaluators. There is one aspect that remains elusive to us. In Spanish, we are not sure why humans see a decrease of quality in terms of `Grammaticality` and `Fluency` on counterfactual data that is also not noticeable in the LLM scores. This will require further investigation to better understand this result.

We looked at system generalisability and robustness through the use of out-of-domain data. The only fully rule-based system submitted (RDFpyreralb) is the most impacted by out-of-domain data, and the least impacted by counterfactual and fictional data. Even though there is little degradation of the LLM-generated texts quality on out-of-domain data, fictional and counterfactual data, it seems like improvements are still achievable on counterfactual and fictional datasets.

The overall interpretation of evaluation results based on mean opinion scores such as the one presented here may be limited, as it is possible that the output quality of the state-of-the-art systems impacts the individual judgments, as noted recently in the speech synthesis domain (Le Maguer et al., 2024). There is an open question for future human evaluations of data-to-text systems on whether a change needs to be made to obtain greater reliability for assessments of intrinsic quality aspects. More generally, the restricted number of systems considered in the analyses, notably for Spanish (three systems), imposes limitations that warrant careful interpretation of the conclusions.

By publicly releasing half of the underlying data (including system outputs, LLM ratings, and human ratings) used to compute the GEM task results, we facilitate further analysis and verification by the research community while preserving portions of the dataset for future experimentation and mitigating potential data leakage. We plan to have a second delayed release and encourage multi-stage data releases given the lack of information about training data.

Acknowledgements

We thank Google for funding our crowdsourcing annotations. Sedoc thanks NYU Stern for their research support. Mille's contribution was funded by the European Union under the Marie Skłodowska-Curie grant agreement No 101062572 (M-FleNS), by the Irish Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media via the eSTÓR project, and Mille benefits from being a member of the SFI Ireland funded ADAPT Research Centre.

References

Alyssa Allen, Ash Lewis, Yi-Chien Lin, Tomiris Kaumenova, and Mike White. 2024. OSU Compling at the GEM'24 data-to-text task. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Agnes Axelsson and Gabriel Skantze. 2023. Using large language models for zero-shot natural language

generation from knowledge graphs. *arXiv preprint arXiv:2307.07312*.

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.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. 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, William Soto-Martinez, et al. 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)*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Kaustubh Dhole, Kai Shu, and Eugene Agichtein. 2025. ConQRet: A new benchmark for fine-grained automatic evaluation of retrieval augmented computational argumentation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5687–5713, Albuquerque, New Mexico. Association for Computational Linguistics.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. 2023. Nl-augmenter: A framework for task-sensitive natural language augmentation.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–27.

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.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori

Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez Beltrachini, Leonardo F . R. Ribeiro, Lewis Tunstall, Li Zhang, Mahim Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

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

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Guy Lapalme. 2024. RDFPYREALB at the GEM'24 data-to-text task: Symbolic english text generation from RDF triples. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Sébastien Le Maguer, Simon King, and Naomi Harte. 2024. The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech & Language*, 84:101577.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Michela Lorandi and Anya Belz. 2024. DCU-NLG-PBN at the GEM'24 data-to-text task: Open-source LLM PEFT-Tuning for effective data-to-text generation. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Simon Mille, Kaustubh D. Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. *ArXiv*, abs/2106.09069.

Simon Mille, Mohammed Sabry, and Anya Belz. 2024a. DCU-NLG-Small at the GEM'24 data-to-text task:

Rule-based generation and post-processing with T5-base. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

Simon Mille, João Sedoc, Yixin Liu, Elizabeth Clark, Agnes Johanna Axelsson, Miruna Adriana Clinciu, Yufang Hou, Saad Mahamood, Ishmael Nyunya Obonyo, and Lining Zhang. 2024b. 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.

Marcel Nawrath, Agnieszka Nowak, Tristan Ratz, Danilo Walenta, Juri Opitz, Leonardo Ribeiro, João Sedoc, Daniel Deutsch, Simon Mille, Yixin Liu, Sebastian Gehrmann, Lining Zhang, Saad Mahamood, Miruna Clinciu, Khyathi Chandu, and Yufang Hou. 2024. On the role of summary content units in text summarization evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 272–281, Mexico City, Mexico. Association for Computational Linguistics.

Michael Oliverio, Pier Felice Balestrucci, Alessandro Mazzei, and Valerio Basile. 2024. DipInfo-UniTo at the GEM'24 data-to-text task: Augmenting LLMs with the split-generate-aggregate pipeline. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Chinonso Cynthia Osuji, Rudali Huidrom, Kolawole John Adebayo, Thiago Castro Ferreira, and Brian Davis. 2024. DCU-ADAPT-modPB at the GEM'24 data-to-text generation task: Model hybridisation for pipeline data-to-text natural language generation. In *Proceedings of the 17th International Conference on Natural Language Generation: Generation Challenges*, Tokyo, Japan. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

John W. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14944–14982, Toronto, Canada. Association for Computational Linguistics.

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

A Prompt for LLM-as-judge

Figure 19 shows the prompt we used for all LLM-as-judge evaluations.

B Complete numerical results tables for English and Spanish

Tables 4 and 5 show all scores obtained by all systems on all datasets.

```
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.
```

Figure 19: 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.

C Details of LLM-as-judge evaluations

The average scores assigned by each LLM to all systems on all datasets is shown in Tables 6 to 9 (English) and Tables 10 to 13 (Spanish).

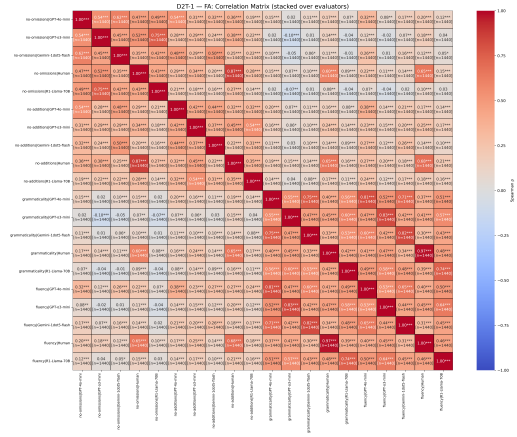
D Details of instance-level correlations on the different datasets

Figures 20 and 21 show the instance-level correlations for each of the 6 datasets in English and Spanish respectively. We computed system–subset correlation matrices to assess the agreement of models across different evaluation criteria. Each input file was identified by a structured filename encoding the *system* (D2T-1 or D2T-2), the *subset* (FA, CFA, or FI), the *evaluator index*, and the *model*. For every file, evaluation columns were first normalized to four canonical dimensions: `No-Omissions`, `No-Additions`, `Grammaticality`, and `Fluency`. Item identifiers were standardized to maximize alignment across files. We then constructed a long-format table in which each row corresponds to a single scored item, annotated with its system, subset, evaluator, model, and criterion. To avoid conflating judgments from different evaluators, the evaluator index was explicitly retained in the item key (i.e. items judged by different evaluators were treated as distinct rows).

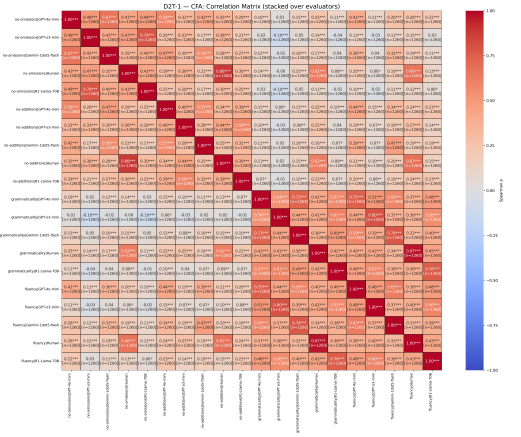
For each system–subset combination, we *stacked* all available evaluators to form a wide-format matrix with rows as items and columns as "criterion–model" pairs. Pairwise Spearman rank correlations (ρ) were then computed between all model–criterion columns using pairwise-complete observations, such that only items scored by both models contributed to a given correlation. Alongside the correlation coefficients, we report two additional statistics: the number of overlapping items used (n), and the two-sided p-value from the Spearman test. The resulting matrices were visualized as annotated heatmaps (six in total, one per system $\times$ subset), where each cell shows ρ, significance markers (* $p < .05$, ** $p < .01$, *** $p < .001$), and n.

E System based plots of LLM vs. Human scores

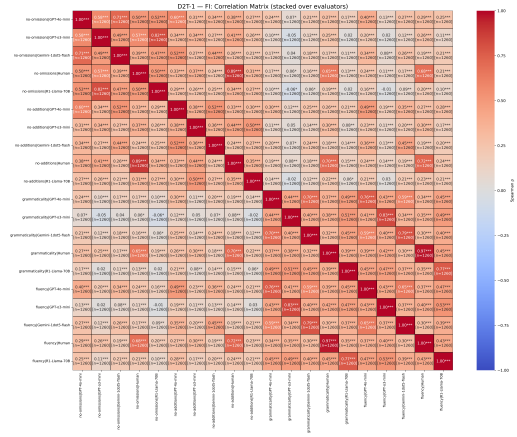
Figures 22 to 25 show plots to visualise the relation between average individual LLM scores (X axis) and average human scores (Y axis) for the English outputs; Figures 26 to 29 show the same for Spanish data.



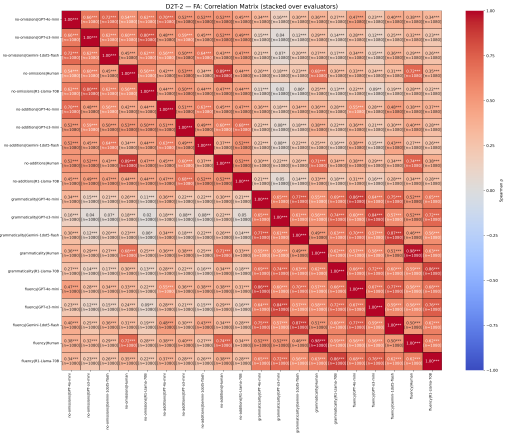
(a) D2T-1 — FA



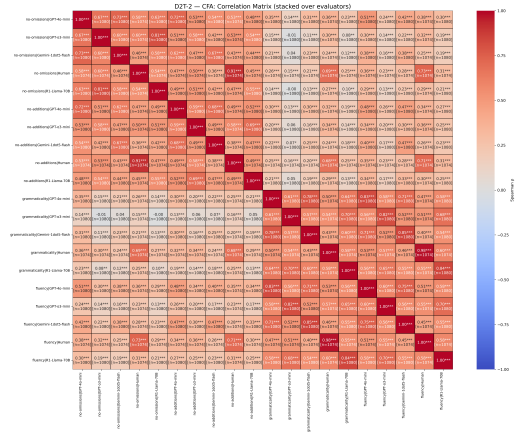
(b) D2T-1 — CFA



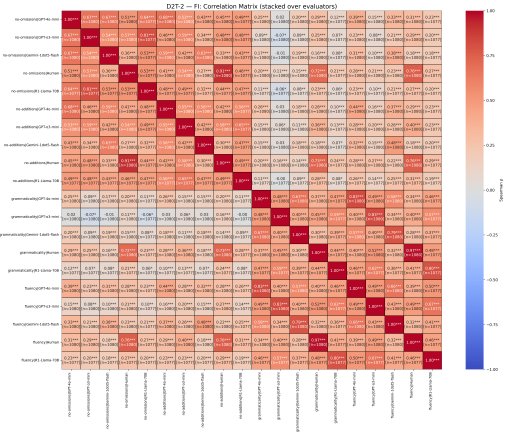
(c) D2T-1 — FI



(d) D2T-2 — FA

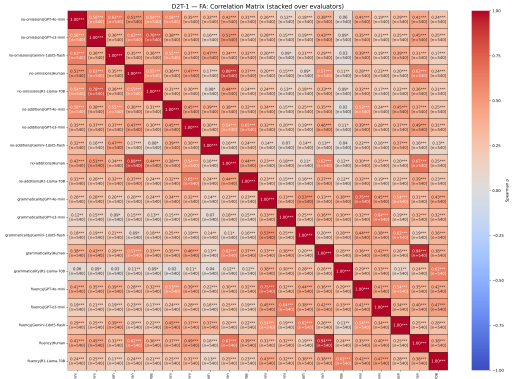


(e) D2T-2 — CFA

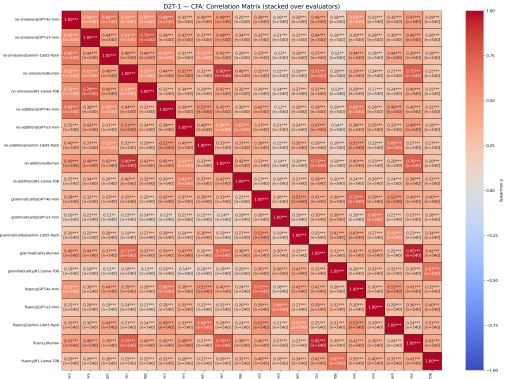


(f) D2T-2 — FI

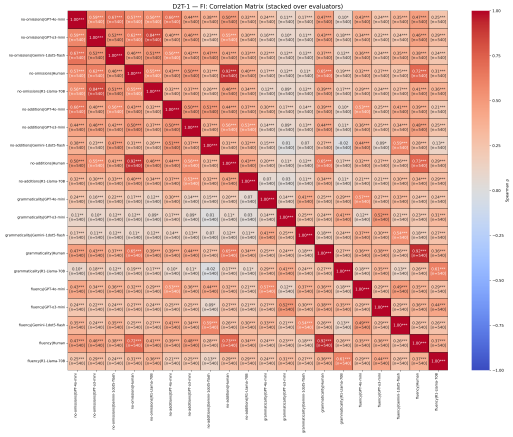
Figure 20: EN System–Subset Correlation Heatmaps: D2T-1 and D2T-2 across FA, CFA, and FI subsets.



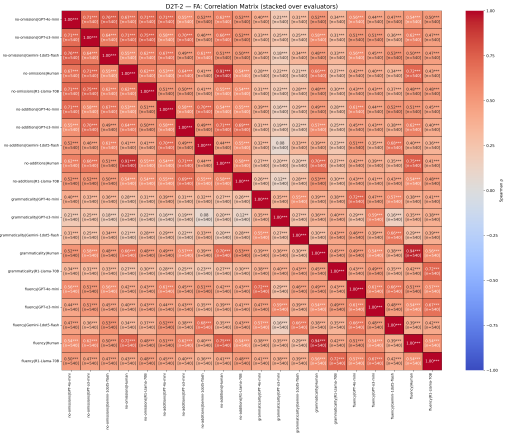
(a) D2T-1 — FA



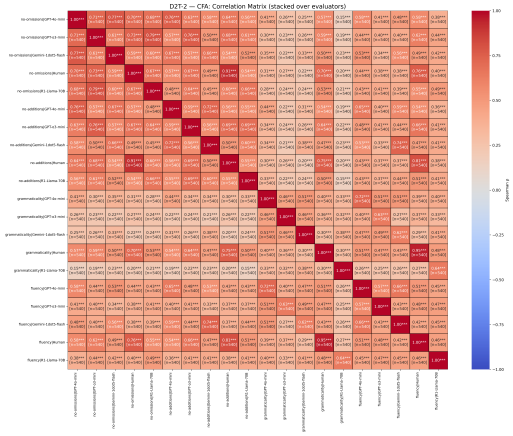
(b) D2T-1 — CFA



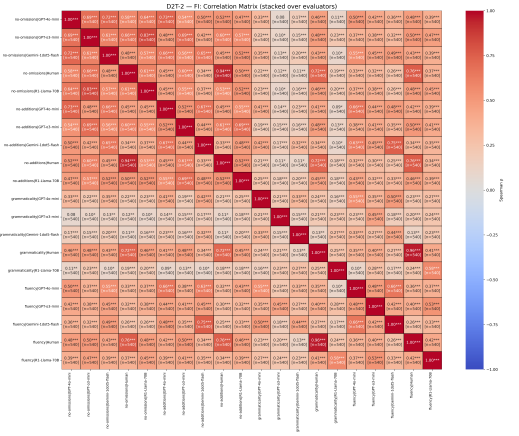
(c) D2T-1 — FI



(d) D2T-2 — FA



(e) D2T-2 — CFA



(f) D2T-2 — FI

Figure 21: ES System–Subset Correlation Heatmaps: D2T-1 and D2T-2 across FA, CFA, and FI subsets.

(EN) Criterion	Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
No-Omissions	Avg. Human ↑	WebNLG-Human	5.14	n/a	n/a	n/a	n/a	n/a	n/a
		DCU-ADAPT-modPB	5.42	5.21	5.35	n/a	n/a	n/a	n/a
		DCU-NLG-PBN	5.49	5.25	5.57	5.46	5.41	5.38	5.43
		DCU-NLG-Small	4.88	4.45	4.46	4.45	4.28	4.3	4.47
		DipInfo-UniTo	5.45	5.43	5.55	5.8	5.72	5.55	5.58
		OSU-CompLing	4.99	4.78	4.54	4.99	4.69	4.4	4.73
		RDFpyrealb	5.74	5.72	5.71	5.46	5.43	5.36	5.57
		SaarLST	5.79	5.52	5.94	6.19	5.93	5.97	5.89
	Avg. LLMs ↑	WebNLG-Human	6.68	n/a	n/a	n/a	n/a	n/a	n/a
		DCU-ADAPT-modPB	6.14	6.19	6.16	n/a	n/a	n/a	n/a
		DCU-NLG-PBN	6.58	6.38	6.65	6.58	6.41	6.73	6.56
		DCU-NLG-Small	6.01	5.6	5.51	5.16	4.96	5.5	5.46
		DipInfo-UniTo	6.51	6.48	6.65	6.65	6.54	6.61	6.57
		OSU-CompLing	6.32	6.13	6.08	6.3	6.15	6.14	6.19
		RDFpyrealb	6.86	6.76	6.82	6.74	6.67	6.82	6.78
		SaarLST	6.86	6.65	6.83	6.97	6.87	6.93	6.85
No-Additions	Avg. Human ↑	WebNLG-Human	5.05	n/a	n/a	n/a	n/a	n/a	n/a
		DCU-ADAPT-modPB	5.82	5.29	5.73	n/a	n/a	n/a	n/a
		DCU-NLG-PBN	5.56	5.1	5.48	5.48	5.08	5.16	5.31
		DCU-NLG-Small	4.85	4.27	4.37	4.42	3.85	4.09	4.31
		DipInfo-UniTo	5.59	5.38	5.47	6.05	5.71	5.39	5.6
		OSU-CompLing	4.85	4.62	4.44	4.97	4.37	4.13	4.56
		RDFpyrealb	5.41	5.41	5.6	5.14	4.96	4.9	5.24
		SaarLST	5.61	5.14	5.76	6.15	5.53	5.76	5.66
	Avg. LLMs ↑	WebNLG-Human	6.67	n/a	n/a	n/a	n/a	n/a	n/a
		DCU-ADAPT-modPB	6.95	6.84	6.94	n/a	n/a	n/a	n/a
		DCU-NLG-PBN	6.88	6.65	6.85	6.7	6.56	6.79	6.74
		DCU-NLG-Small	6.42	6.25	6.1	5.5	5.33	5.91	5.92
		DipInfo-UniTo	6.88	6.78	6.86	6.89	6.82	6.77	6.83
		OSU-CompLing	6.68	6.57	6.6	6.69	6.58	6.45	6.6
		RDFpyrealb	6.86	6.76	6.84	6.6	6.59	6.74	6.73
		SaarLST	6.83	6.53	6.88	6.89	6.79	6.89	6.8
Grammaticality	Avg. Human ↑	WebNLG-Human	5.43	n/a	n/a	n/a	n/a	n/a	n/a
		DCU-ADAPT-modPB	6.39	6.08	6.18	n/a	n/a	n/a	n/a
		DCU-NLG-PBN	6.11	5.68	5.86	6.01	5.67	5.44	5.79
		DCU-NLG-Small	5.51	5.12	5.26	5.01	4.7	4.96	5.09
		DipInfo-UniTo	6.01	5.68	5.81	6.12	5.95	5.55	5.85
		OSU-CompLing	5.59	5.02	5.03	5.49	4.93	4.84	5.15
		RDFpyrealb	4.53	4.66	4.89	4.1	4.11	4.34	4.44
		SaarLST	6.07	5.83	5.98	6.28	6.08	6.01	6.04
	Avg. LLMs ↑	WebNLG-Human	6.77	n/a	n/a	n/a	n/a	n/a	n/a
		DCU-ADAPT-modPB	6.99	6.97	6.99	n/a	n/a	n/a	n/a
		DCU-NLG-PBN	6.99	6.91	6.93	6.94	6.96	6.97	6.95
		DCU-NLG-Small	6.87	6.81	6.82	6.6	6.58	6.83	6.75
		DipInfo-UniTo	6.96	6.94	6.97	6.89	6.91	6.86	6.92
		OSU-CompLing	6.82	6.65	6.76	6.76	6.68	6.84	6.75
		RDFpyrealb	6.13	6.04	6.26	5.38	5.45	6.03	5.88
		SaarLST	6.95	6.94	6.97	6.98	6.98	6.99	6.97
Fluency	Avg. Human ↑	WebNLG-Human	5.41	n/a	n/a	n/a	n/a	n/a	n/a
		DCU-ADAPT-modPB	6.29	5.97	6.1	n/a	n/a	n/a	n/a
		DCU-NLG-PBN	6.04	5.6	5.81	5.92	5.63	5.46	5.74
		DCU-NLG-Small	5.5	5.01	5.23	4.99	4.74	4.94	5.07
		DipInfo-UniTo	5.89	5.58	5.72	6.06	5.9	5.53	5.78
		OSU-CompLing	5.61	5.1	5.16	5.55	5.02	4.91	5.23
		RDFpyrealb	4.69	4.75	4.99	4.35	4.29	4.54	4.6
		SaarLST	5.98	5.76	5.94	6.24	6.0	5.95	5.98
	Avg. LLMs ↑	WebNLG-Human	6.75	n/a	n/a	n/a	n/a	n/a	n/a
		DCU-ADAPT-modPB	6.99	6.95	6.98	n/a	n/a	n/a	n/a
		DCU-NLG-PBN	6.96	6.88	6.93	6.93	6.92	6.97	6.93
		DCU-NLG-Small	6.8	6.68	6.71	6.42	6.39	6.65	6.61
		DipInfo-UniTo	6.92	6.87	6.94	6.84	6.87	6.8	6.87
		OSU-CompLing	6.86	6.74	6.83	6.86	6.75	6.88	6.82
		RDFpyrealb	6.1	5.98	6.25	5.34	5.42	6.01	5.85
		SaarLST	6.94	6.9	6.95	6.98	6.96	6.98	6.95

Table 4: Qualitative scores for the English D2T task (180 data points).

(ES) Criterion	Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
No-Omissions	**Avg. Human ↑**	DCU-NLG-PBN	5.96	5.76	5.93	5.94	5.82	5.81	5.87
		DCU-NLG-Small	5.12	5.07	4.61	4.55	4.3	4.62	4.71
		OSU-CompLing	6.18	6.0	6.09	6.03	5.98	6.13	6.07
	Avg. LLMs ↑	DCU-NLG-PBN	6.54	6.39	6.58	6.61	6.41	6.73	6.54
		DCU-NLG-Small	6.02	5.6	5.62	5.2	5.0	5.58	5.5
		OSU-CompLing	6.78	6.66	6.77	6.79	6.63	6.82	6.74
No-Additions	**Avg. Human ↑**	DCU-NLG-PBN	5.91	5.47	5.81	5.84	5.33	5.51	5.65
		DCU-NLG-Small	5.02	4.66	4.57	4.34	3.79	4.25	4.44
		OSU-CompLing	5.89	5.58	5.9	5.68	5.42	5.77	5.71
	Avg. LLMs ↑	DCU-NLG-PBN	6.9	6.7	6.88	6.72	6.58	6.82	6.77
		DCU-NLG-Small	6.48	6.26	6.2	5.67	5.46	6.0	6.01
		OSU-CompLing	6.77	6.67	6.8	6.74	6.68	6.78	6.74
Grammaticality	**Avg. Human ↑**	DCU-NLG-PBN	6.72	6.39	6.58	6.67	6.6	6.47	6.57
		DCU-NLG-Small	6.12	5.91	6.0	5.58	5.38	5.65	5.77
		OSU-CompLing	6.72	6.53	6.73	6.67	6.51	6.54	6.61
	Avg. LLMs ↑	DCU-NLG-PBN	6.98	6.93	6.96	6.97	6.95	6.97	6.96
		DCU-NLG-Small	6.82	6.77	6.89	6.71	6.71	6.89	6.8
		OSU-CompLing	6.97	6.96	6.98	6.98	6.97	6.98	6.97
Fluency	**Avg. Human ↑**	DCU-NLG-PBN	6.68	6.31	6.55	6.64	6.54	6.45	6.53
		DCU-NLG-Small	6.06	5.83	5.96	5.54	5.32	5.63	5.72
		OSU-CompLing	6.7	6.45	6.69	6.63	6.49	6.53	6.58
	Avg. LLMs ↑	DCU-NLG-PBN	6.97	6.93	6.97	6.96	6.96	6.99	6.96
		DCU-NLG-Small	6.77	6.67	6.81	6.54	6.55	6.75	6.68
		OSU-CompLing	6.97	6.95	6.98	6.97	6.95	6.97	6.97

Table 5: Qualitative scores for the Spanish D2T task (180 data points).

Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
GPT-4o-mini ↑	WebNLG-Human	6.38	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.09	6.06	6.14	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.43	6.16	6.64	6.47	6.26	6.66	6.44
	DCU-NLG-Small	5.84	5.37	5.36	5.1	4.7	5.42	5.3
	DipInfo-UniTo	6.33	6.32	6.63	6.63	6.38	6.58	6.48
	OSU-CompLing	6.14	5.83	5.89	6.14	5.86	5.94	5.97
	RDFpyrealb	6.62	6.44	6.58	6.41	6.21	6.65	6.49
	SaarLST	6.72	6.38	6.81	6.94	6.78	6.91	6.76
o3-mini ↑	WebNLG-Human	6.65	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.04	6.13	6.06	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.52	6.37	6.52	6.54	6.43	6.66	6.51
	DCU-NLG-Small	5.78	5.55	5.29	4.88	4.77	5.18	5.24
	DipInfo-UniTo	6.37	6.46	6.57	6.58	6.54	6.52	6.51
	OSU-CompLing	6.17	6.05	5.93	6.22	6.2	6.03	6.1
	RDFpyrealb	6.95	6.97	6.93	6.81	6.85	6.84	6.89
	SaarLST	6.88	6.77	6.78	6.96	6.91	6.94	6.87
Gemini-1.5-flash ↑	WebNLG-Human	6.82	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.42	6.37	6.36	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.8	6.56	6.84	6.76	6.42	6.88	6.71
	DCU-NLG-Small	6.33	5.92	5.96	5.55	5.31	6.02	5.85
	DipInfo-UniTo	6.76	6.66	6.78	6.74	6.59	6.74	6.71
	OSU-CompLing	6.58	6.51	6.53	6.57	6.42	6.52	6.52
	RDFpyrealb	6.9	6.72	6.82	6.83	6.73	6.94	6.82
	SaarLST	6.97	6.68	6.92	7.0	6.86	6.96	6.9
R1-Llama-70B ↑	WebNLG-Human	6.85	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	5.99	6.21	6.08	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.57	6.43	6.6	6.57	6.53	6.73	6.57
	DCU-NLG-Small	6.09	5.57	5.44	5.09	5.04	5.37^{i}	5.44^{i}
	DipInfo-UniTo	6.57	6.46	6.64	6.63	6.64	6.6^{i}	6.59^{i}
	OSU-CompLing	6.39	6.13	5.96	6.27	6.13	6.06	6.16
	RDFpyrealb	6.97	6.92	6.94	6.91	6.88	6.86	6.91
	SaarLST	6.88	6.79	6.83	6.97	6.93	6.92^{i}	6.89^{i}

Table 6: LLM-as-judge scores for `No-Omissions` on the English D2T task. Each score in the table is the average of 180 scores, except when indicated otherwise: ^{i} one score missing.

Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
GPT-4o-mini ↑	WebNLG-Human	6.72	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.94	6.81	6.92	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.92	6.74	6.92	6.76	6.66	6.88	6.81
	DCU-NLG-Small	6.53	6.24	6.11	5.72	5.39	5.98	6.0
	DipInfo-UniTo	6.83	6.76	6.91	6.91	6.76	6.74	6.82
	OSU-CompLing	6.71	6.59	6.55	6.84	6.62	6.43	6.62
	RDFpyrealb	6.86	6.67	6.76	6.54	6.52	6.77	6.69
	SaarLST	6.87	6.43	6.88	6.93	6.87	6.88	6.81
o3-mini ↑	WebNLG-Human	6.41	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.93	6.87	6.91	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.78	6.45	6.68	6.47	6.3	6.58	6.54
	DCU-NLG-Small	6.07	5.98	5.58	5.04	4.92	5.49	5.51
	DipInfo-UniTo	6.81	6.66	6.72	6.84	6.79	6.62	6.74
	OSU-CompLing	6.39	6.37	6.36	6.34	6.34	6.14	6.32
	RDFpyrealb	6.76	6.79	6.76	6.4	6.39	6.46	6.59
	SaarLST	6.64	6.39	6.77	6.77	6.57	6.83	6.66
Gemini-1.5-flash ↑	WebNLG-Human	6.92	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.99	6.85	6.99	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.98	6.82	6.95	6.95	6.73	6.91	6.89
	DCU-NLG-Small	6.71	6.5	6.55	5.96	5.67	6.35	6.29
	DipInfo-UniTo	6.97	6.93	6.98	6.94	6.88	6.94	6.94
	OSU-CompLing	6.91	6.8	6.92	6.92	6.81	6.83	6.86
	RDFpyrealb	6.94	6.85	6.96	6.84	6.82	6.95	6.89
	SaarLST	6.97	6.79	7.0	6.99	6.96	6.93	6.94
R1-Llama-70B ↑	WebNLG-Human	6.66	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.92	6.83	6.95	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.84	6.59	6.87	6.64	6.54	6.78	6.71
	DCU-NLG-Small	6.37	6.28	6.15	5.27	5.32	5.82^{i}	5.87^{i}
	DipInfo-UniTo	6.89	6.78	6.85	6.88	6.83	6.78^{i}	6.84^{i}
	OSU-CompLing	6.73	6.53	6.59	6.66	6.56	6.39	6.58
	RDFpyrealb	6.88	6.74	6.88	6.63	6.63	6.79	6.76
	SaarLST	6.83	6.51	6.88	6.86	6.78	6.93^{i}	6.8^{i}

Table 7: LLM-as-judge scores for `No-Additions` on the English D2T task. Each score in the table is the average of 180 scores, except when indicated otherwise: ^{i} one score missing.

Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
GPT-4o-mini ↑	WebNLG-Human	6.83	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	7.0	6.99	7.0	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.99	6.95	6.98	6.96	6.97	7.0	6.98
	DCU-NLG-Small	6.85	6.84	6.83	6.62	6.59	6.82	6.76
	DipInfo-UniTo	6.97	6.97	7.0	6.91	6.97	6.89	6.95
	OSU-CompLing	6.84	6.65	6.74	6.77	6.71	6.85	6.76
	RDFpyrealb	6.35	6.29	6.48	5.73	5.79	6.46	6.18
	SaarLST	6.98	6.98	6.97	6.99	7.0	7.0	6.99
o3-mini ↑	WebNLG-Human	6.62	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	7.0	6.98	6.99	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.98	6.91	6.91	6.97	7.0	6.94	6.95
	DCU-NLG-Small	6.92	6.87	6.83	6.69	6.72	6.91	6.82
	DipInfo-UniTo	6.95	6.98	6.98	6.84	6.85	6.81	6.9
	OSU-CompLing	6.81	6.73	6.85	6.79	6.71	6.87	6.79
	RDFpyrealb	5.56	5.38	5.62	4.72	4.73	5.16	5.19
	SaarLST	6.92	6.89	6.98	6.97	6.98	6.98	6.95
Gemini-1.5-flash ↑	WebNLG-Human	6.92	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	7.0	6.99	7.0	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	7.0	6.98	6.98	6.97	6.99	7.0	6.99
	DCU-NLG-Small	6.9	6.84	6.88	6.7	6.63	6.88	6.81
	DipInfo-UniTo	6.99	6.97	7.0	6.96	6.96	6.94	6.97
	OSU-CompLing	6.85	6.73	6.84	6.86	6.78	6.96	6.84
	RDFpyrealb	6.49	6.53	6.67	5.92	6.08	6.63	6.39
	SaarLST	7.0	7.0	6.99	6.99	6.99	7.0	6.99
R1-Llama-70B ↑	WebNLG-Human	6.7	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.97	6.93	6.97	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.97	6.81	6.86	6.86	6.87	6.94	6.89
	DCU-NLG-Small	6.81	6.69	6.76	6.38	6.39	6.71^{i}	6.62^{i}
	DipInfo-UniTo	6.95	6.83	6.89	6.84	6.85	6.79^{i}	6.86^{i}
	OSU-CompLing	6.76	6.47	6.62	6.62	6.52	6.69	6.61
	RDFpyrealb	6.14	5.95	6.26	5.13	5.21	5.88	5.76
	SaarLST	6.91	6.9	6.94	6.97	6.97	6.98^{i}	6.94^{i}

Table 8: LLM-as-judge scores for `Grammaticality` on the English D2T task. Each score in the table is the average of 180 scores, except when indicated otherwise: ^{i} one score missing.

Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
GPT-4o-mini ↑	WebNLG-Human	6.79	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.98	6.95	6.98	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.97	6.89	6.98	6.93	6.92	7.0	6.95
	DCU-NLG-Small	6.78	6.66	6.69	6.41	6.36	6.65	6.59
	DipInfo-UniTo	6.93	6.9	6.97	6.87	6.92	6.86	6.91
	OSU-CompLing	6.85	6.72	6.78	6.87	6.72	6.88	6.8
	RDFpyrealb	6.36	6.22	6.46	5.78	5.8	6.42	6.17
	SaarLST	6.97	6.9	6.96	6.99	6.98	7.0	6.97
o3-mini ↑	WebNLG-Human	6.66	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	7.0	6.98	6.99	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.97	6.9	6.93	6.96	6.96	6.95	6.94
	DCU-NLG-Small	6.86	6.72	6.73	6.49	6.49	6.68	6.66
	DipInfo-UniTo	6.92	6.96	6.96	6.83	6.84	6.77	6.88
	OSU-CompLing	6.92	6.84	6.94	6.92	6.85	6.91	6.9
	RDFpyrealb	5.67	5.6	5.86	4.85	4.9	5.36	5.37
	SaarLST	6.92	6.93	6.97	6.99	6.99	6.98	6.96
Gemini-1.5-flash ↑	WebNLG-Human	6.94	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	7.0	6.97	7.0	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	7.0	6.96	6.98	6.97	6.98	6.99	6.98
	DCU-NLG-Small	6.89	6.79	6.83	6.62	6.53	6.8	6.75
	DipInfo-UniTo	6.98	6.96	7.0	6.91	6.93	6.92	6.95
	OSU-CompLing	6.93	6.87	6.92	6.93	6.89	6.98	6.92
	RDFpyrealb	6.57	6.55	6.68	6.05	6.14	6.65	6.44
	SaarLST	6.99	6.98	6.99	7.0	6.99	7.0	6.99
R1-Llama-70B ↑	WebNLG-Human	6.61	n/a	n/a	n/a	n/a	n/a	n/a
	DCU-ADAPT-modPB	6.97	6.9	6.96	n/a	n/a	n/a	n/a
	DCU-NLG-PBN	6.92	6.76	6.84	6.86	6.83	6.93	6.86
	DCU-NLG-Small	6.66	6.53	6.57	6.17	6.19	6.46^{i}	6.43^{i}
	DipInfo-UniTo	6.84	6.68	6.82	6.78	6.79	6.65^{i}	6.76^{i}
	OSU-CompLing	6.75	6.54	6.68	6.72	6.52	6.74	6.66
	RDFpyrealb	5.79	5.56	5.98	4.69	4.83	5.59	5.41
	SaarLST	6.89	6.77	6.89	6.94	6.87	6.93^{i}	6.88^{i}

Table 9: LLM-as-judge scores for `Fluency` on the English D2T task. Each score in the table is the average of 180 scores, except when indicated otherwise: ^{i} one score missing.

Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
GPT-4o-mini ↑	DCU-NLG-PBN	6.29	6.11	6.49	6.46	6.18	6.64	6.36
	DCU-NLG-Small	5.66	5.26	5.43	4.87	4.65	5.5	5.23
	OSU-CompLing	6.71	6.45	6.72	6.67	6.37	6.82	6.62
o3-mini ↑	DCU-NLG-PBN	6.62	6.42	6.53	6.59	6.44	6.67	6.54
	DCU-NLG-Small	5.82	5.57	5.41	4.99	4.78	5.24	5.3
	OSU-CompLing	6.75	6.67	6.72	6.72	6.71	6.71	6.71
Gemini-1.5-flash ↑	DCU-NLG-PBN	6.61	6.54	6.74	6.68	6.44	6.86	6.65
	DCU-NLG-Small	6.29	5.87	5.98	5.38	5.24	5.98	5.79
	OSU-CompLing	6.84	6.67	6.81	6.88	6.63	6.89	6.79
R1-Llama-70B ↑	DCU-NLG-PBN	6.64	6.48	6.55	6.72	6.57	6.73	6.62
	DCU-NLG-Small	6.32	5.68	5.66	5.55	5.32	5.6	5.69
	OSU-CompLing	6.83	6.83	6.84	6.88	6.79	6.86	6.84

Table 10: LLM-as-judge scores for `No-Omissions` on the Spanish D2T task. Each score in the table is the average of 180 scores.

Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
GPT-4o-mini ↑	DCU-NLG-PBN	6.83	6.65	6.91	6.76	6.58	6.84	6.76
	DCU-NLG-Small	6.57	6.23	6.24	5.71	5.49	6.14	6.06
	OSU-CompLing	6.83	6.77	6.87	6.78	6.73	6.84	6.8
o3-mini ↑	DCU-NLG-PBN	6.86	6.58	6.78	6.5	6.34	6.63	6.62
	DCU-NLG-Small	6.07	5.96	5.72	5.08	4.96	5.42	5.53
	OSU-CompLing	6.59	6.38	6.6	6.41	6.54	6.6	6.52
Gemini-1.5-flash ↑	DCU-NLG-PBN	6.99	6.89	6.96	6.92	6.81	6.97	6.92
	DCU-NLG-Small	6.8	6.55	6.65	6.11	5.74	6.44	6.38
	OSU-CompLing	6.93	6.77	6.94	6.97	6.72	6.92	6.87
R1-Llama-70B ↑	DCU-NLG-PBN	6.93	6.66	6.88	6.72	6.59	6.84	6.77
	DCU-NLG-Small	6.49	6.29	6.18	5.77	5.64	5.98	6.06
	OSU-CompLing	6.73	6.74	6.78	6.81	6.74	6.78	6.76

Table 11: LLM-as-judge scores for `No-Additions` on the Spanish D2T task. Each score in the table is the average of 180 scores.

Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
GPT-4o-mini ↑	DCU-NLG-PBN	6.99	6.96	6.98	6.98	6.94	6.99	6.97
	DCU-NLG-Small	6.79	6.75	6.89	6.69	6.71	6.89	6.79
	OSU-CompLing	7.0	6.96	6.99	6.98	6.97	6.99	6.98
o3-mini ↑	DCU-NLG-PBN	6.94	6.88	6.95	6.96	6.96	6.97	6.94
	DCU-NLG-Small	6.77	6.77	6.87	6.77	6.7	6.92	6.8
	OSU-CompLing	6.91	6.94	6.96	6.96	6.94	6.98	6.95
Gemini-1.5-flash ↑	DCU-NLG-PBN	7.0	7.0	6.99	6.99	6.99	7.0	7.0
	DCU-NLG-Small	6.9	6.87	6.94	6.81	6.79	6.95	6.88
	OSU-CompLing	7.0	6.99	7.0	7.0	7.0	7.0	7.0
R1-Llama-70B ↑	DCU-NLG-PBN	6.98	6.91	6.91	6.94	6.93	6.93	6.93
	DCU-NLG-Small	6.79	6.71	6.86	6.56	6.65	6.82	6.73
	OSU-CompLing	6.97	6.95	6.96	6.96	6.96	6.96	6.96

Table 12: LLM-as-judge scores for `Grammaticality` on the Spanish D2T task. Each score in the table is the average of 180 scores.

Evaluator	System	D2T-1 FA	D2T-1 CFA	D2T-1 FI	D2T-2 FA	D2T-2 CFA	D2T-2 FI	Avg.
GPT-4o-mini ↑	DCU-NLG-PBN	6.96	6.91	6.98	6.98	6.94	6.99	6.96
	DCU-NLG-Small	6.75	6.62	6.82	6.5	6.48	6.77	6.66
	OSU-CompLing	6.99	6.94	6.99	6.97	6.94	6.98	6.97
o3-mini ↑	DCU-NLG-PBN	6.96	6.94	6.97	6.97	6.99	6.99	6.97
	DCU-NLG-Small	6.73	6.67	6.79	6.51	6.56	6.69	6.66
	OSU-CompLing	6.93	6.93	6.98	6.94	6.94	6.98	6.95
Gemini-1.5-flash ↑	DCU-NLG-PBN	6.99	6.98	7.0	6.98	6.97	7.0	6.99
	DCU-NLG-Small	6.9	6.81	6.9	6.74	6.68	6.88	6.82
	OSU-CompLing	7.0	6.98	6.99	6.99	6.97	6.99	6.99
R1-Llama-70B ↑	DCU-NLG-PBN	6.97	6.88	6.92	6.93	6.95	6.96	6.94
	DCU-NLG-Small	6.71	6.58	6.74	6.42	6.48	6.66	6.6
	OSU-CompLing	6.96	6.93	6.97	6.97	6.96	6.94	6.95

Table 13: LLM-as-judge scores for `Fluency` on the Spanish D2T task. Each score in the table is the average of 180 scores.

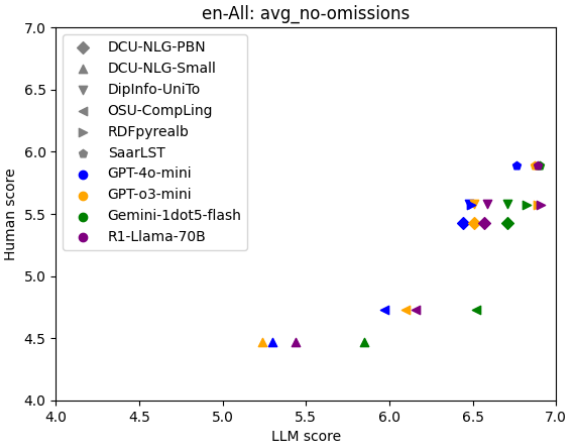


Figure 22: Plot of LLM scores against Human scores for all systems: English, `No-Omissions`

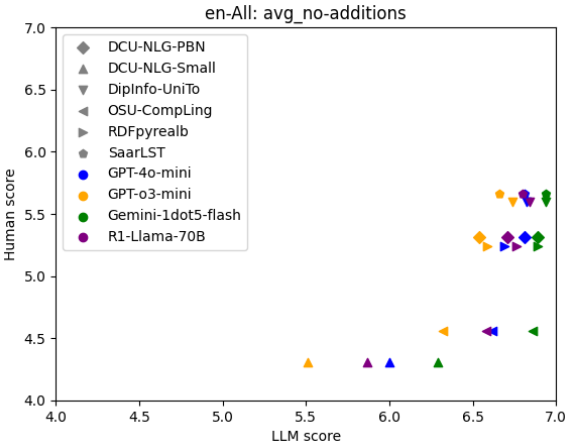


Figure 23: Plot of LLM scores against Human scores for all systems: English, `No-Additions`

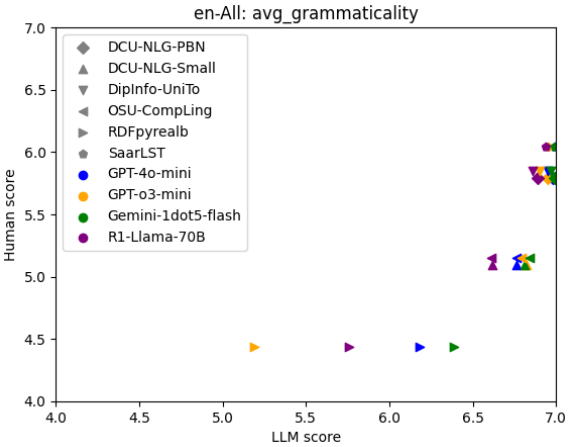


Figure 24: Plot of LLM scores against Human scores for all systems: English, `Grammaticality`

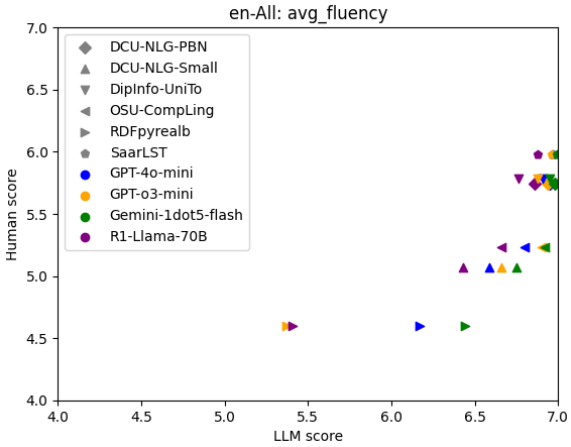


Figure 25: Plot of LLM scores against Human scores for all systems: English, `Fluency`

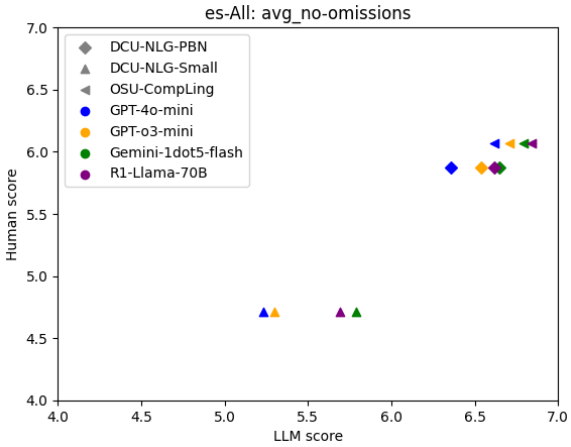


Figure 26: Plot of LLM scores against Human scores for all systems: Spanish, `No-Omissions`

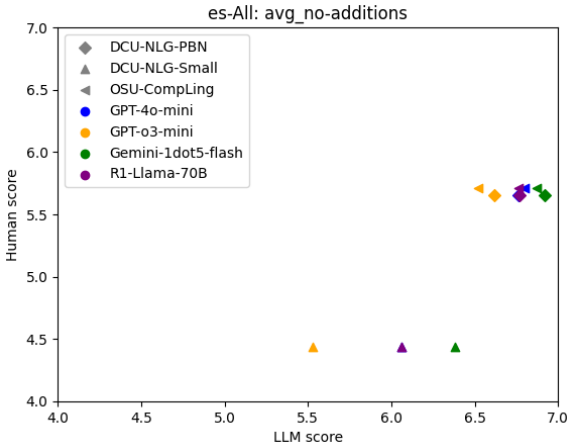


Figure 27: Plot of LLM scores against Human scores for all systems: Spanish, `No-Additions`

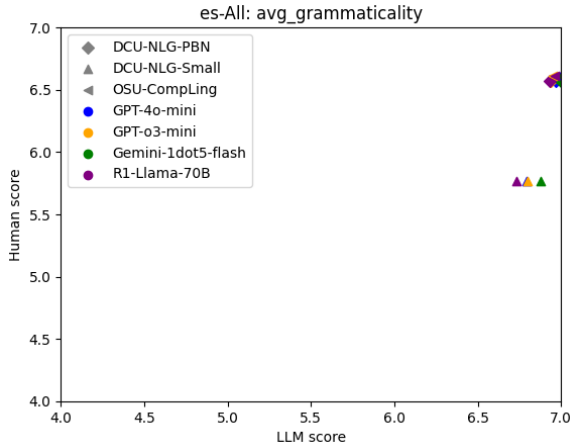


Figure 28: Plot of LLM scores against Human scores for all systems: Spanish, `Grammaticality`

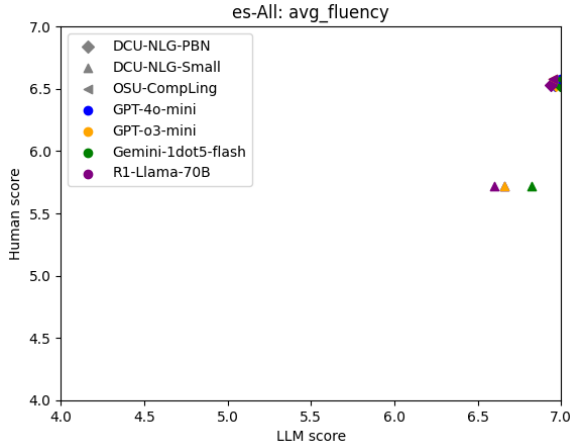


Figure 29: Plot of LLM scores against Human scores for all systems: Spanish, `Fluency`

Live Commentary Planning and Generation

Chung-Chi Chen,[1] **Huan-Wen Ho,**[2] **Yu-Yu Chang,**[2] **Ming-Hung Wang,**[2]
Ramon Ruiz-Dolz,[3] **Chris Reed**,[3] **Ichiro Kobayashi**,[4] **Yusuke Miyao**,[5] **Hiroya Takamura**[1]
[1]AIST, Japan
[2]National Chung Cheng University, Taiwan
[3]Centre for Argument Technology (ARG-tech), University of Dundee, UK
[4]Ochanomizu University, Japan
[5]University of Tokyo, Japan

Abstract

Live commentary plays a crucial role in helping audiences interpret high-stakes events such as political debates, central bank press conferences, and corporate earnings calls. Unlike generic summarization, professional commentary requires timely decisions about *what* to comment on and *how* to present it, integrating fact-checking, background knowledge, and subjective evaluation. However, little prior work has studied commentary as a structured planning and generation problem. To bridge this gap, we introduce the first multi-domain dataset of **Live Commentary Planning and Generation**, aligning event transcripts with time-synchronized expert analyses and public reactions. Our dataset covers U.S. presidential debates (2016–2024), Federal Open Market Committee press conferences, and corporate earnings calls, enriched with a fine-grained taxonomy of commentary intents (up to 11 categories) and supplemented by Reddit crowd commentary. We define two benchmark tasks: (1) *Commentary Planning*, predicting the type of commentary given a transcript segment, and (2) *Commentary Generation*, producing commentary text conditioned on the segment and a target label. Baseline experiments with large language models show that, despite their fluency, models struggle with expert-level commentary, showing the difficulty of integrating contextual reasoning and external knowledge under real-time constraints.[1]

1 Introduction

Large language models (LLMs) have made it easier than ever to generate fluent text, but true professional-quality commentary demands more than fluency. In high-stakes public discourse, such as political debates, central bank press conferences, or corporate earnings calls, expert commentators provide real-time analysis that contextualizes and critiques what is being said. This live commentary helps transform passive viewership into an engaged, informed experience. Professional commentators translate complex language into accessible insights, fact-check claims in real-time, and offer historical or expert perspective to guide audience understanding. Such commentary must be timely, knowledgeable, and context-rich, going beyond summarization to include opinions, fact-checks, and interpretations. However, simulating this expert ability is challenging: it requires deciding what to comment on (planning) and how to convey it (generation) under time pressure and with domain expertise.

Despite extensive research on these domains individually (e.g. analyzing debate transcripts or summarizing financial reports), little work has aligned transcripts with their simultaneous expert commentary. Existing studies tend to treat the primary content and the reactions separately. For example, focusing on debate speeches or on social media responses in isolation. This leaves a gap in understanding how experts interpret dynamic events in the moment. To address this gap, we introduce a new dataset of Live Commentary Planning and Generation that aligns real-time expert commentary with transcripts across three domains: (1) U.S. presidential debates (2016–2024), (2) Federal Open Market Committee (FOMC) press conferences, and (3) corporate earnings calls. In each setting, multiple expert commentators observed the live event and produced running commentary, which we have collected and aligned with the spoken transcript segments by timestamp. Additionally, for the presidential debates, we incorporate public commentary from Reddit discussion threads to capture non-expert, crowd reactions in real time. The result is a multi-faceted dataset covering both institutional expert analysis and grassroots public reactions.

Crucially, each commentary segment is annotated with a fine-grained category from a new taxonomy we developed for live discourse analysis.

[1]Project Page: http://livecommentary.nlpfin.com/

Proceedings of the 18th International Natural Language Generation Conference: Generation Challenges, pages 37–43
October 29–November 2, 2025.

For example, debate commentary segments are labeled as Key Summary, Supplementary Explanation, Fact-Check, Personal Opinion, Market Reaction, Public Opinion, or Commentator's Question, with Personal Opinion further broken into subtypes like evaluating performance, analyzing claims, drawing inferences, etc.. This rich labeling (11 distinct labels in total for debates) enables models to learn not just to generate commentary, but to plan what type of comment is appropriate at each moment. In professional settings, the ability to choose an apt perspective, e.g. to fact-check a dubious claim versus to summarize a complex point, is critical. Our dataset supports two complementary tasks: (1) Commentary Planning, i.e. predicting the commentary label given a transcript segment, and (2) Commentary Generation, i.e. producing the content of a commentary given the segment and a target label. By tackling these tasks, models must learn to mimic expert decision-making and contextual writing under real-time constraints.

In summary, our contributions are: (a) a first-of-its-kind dataset aligning transcripts with real-time expert commentary across multiple domains, with fine-grained annotations of commentary intent; (b) benchmark task definitions for commentary planning and generation, to facilitate systematic study of this challenging form of conditional text generation; and (c) initial analyses and baseline results demonstrating the dataset's difficulty and the need for advanced techniques. Even state-of-the-art LLMs like GPT-4 struggle with expert commentary planning and generation (as shown in our pilot studies), underscoring the novelty and challenge of our task. We hope this dataset will spur research at the intersection of content understanding, knowledge integration, and real-time text generation.

2 Dataset

2.1 Dataset Creation

Our dataset encompasses three types of live events: (1) U.S. presidential debates, (2) FOMC press conferences, and (3) corporate earnings calls , along with their live commentary. For U.S. presidential debates, we include all major televised debates from the 2016, 2020, and 2024 election cycles. This totals 10 events (including presidential and vice-presidential debates and a 2023 primary debate), with full transcripts obtained from public sources (e.g. debate commission or media outlets). We collected the real-time expert commentary on these debates from the Bloomberg news service, which had professional journalists providing line-by-line analysis during the live broadcasts. Each commentary piece is timestamped. We align each commentary segment to the corresponding part of the debate transcript by timestamp and content, ensuring the commentator's remark is matched with the specific speaker utterance or segment it addresses. If a comment does not clearly relate to any specific line, it is marked as not applicable to a segment. Using this procedure, we obtained 2,283 commentary-transcript pairs for debates.

	Debates	FOMC	Earnings Call	Reddit
# Pair	2,283	252	1,115	366
# Category	11	5	10	4

Table 1: Dataset statistics.

For FOMC press conferences, we collected transcripts of the Fed Chair's opening statement and the subsequent Q&A with journalists, for multiple meetings, covering 8 FOMC events. We again used Bloomberg's real-time commentary feed, which provides expert economist reactions during these press conferences. After alignment, we have 252 commentary segments paired with FOMC transcript segments. For corporate earnings calls, we focus on earnings calls of S&P 500 companies across various sectors. Earnings calls typically consist of a management presentation and a Q&A session with analysts. We use transcripts and align Bloomberg's live financial commentary on those calls. The dataset includes 1,115 pairs of commentary with earnings call transcript segments. Lastly, for Reddit commentary, we incorporate public reactions from Reddit "mega-threads" created during the 2016 U.S. presidential debates. Using an Intertextual Topic Correspondence (ITC) method (Visser et al., 2018), we matched 366 Reddit comments to relevant debate utterances. These alignments were verified and annotated with simplified labels (described below). The inclusion of Reddit allows us to compare expert vs. crowd commentary directly.

2.2 Label Taxonomy

Table 1 summarizes the size of each domain in our dataset and the label inventory available. In total, the dataset contains over 3,650 expert commentary instances aligned with transcripts (plus 366 Reddit instances), making it the largest resource of its kind to date.

Label	Description
KS (Key Summary)	Summarizing what the speaker said.
SE (Supplementary Explanation)	Providing additional factual context or background (often drawing on external knowledge).
FC (Fact-Checking)	Verifying or refuting the accuracy of a candidate's claim.
PO (Public Opinion)	Noting public sentiment or likely voter reactions (sometimes referencing polls or social media).
MR (Market Reaction)	Commenting on any immediate financial market response or economic implications (included since commentators are financial journalists).
CQ (Commentator's Question)	Posing an open question or something to watch for (e.g., "How will candidate X implement this policy?").
CPO (Commentator's Personal Opinion)	Any subjective analysis or evaluative remark by the commentator. Expanded into five finer labels:
PC (Performance Critique)	Evaluating the debate performance or rhetorical style of the participants.
CS (Claim Analysis)	Opining on specific policy claims or factual statements made.
AC (Analytical Conclusion)	Drawing a conclusion or inference beyond the given facts.
MP (Market/Policy Projection)	Connecting the debate content to economic or policy outcomes (e.g., impact on markets).
O (Other)	Any opinion-based comment that doesn't fit the above (catch-all).

Table 2: Commentary labels in debates

Each domain has a tailored commentary taxonomy reflecting the nature of that discourse, while maintaining some common themes. For the debates, we developed a hierarchical label schema with 7 main categories and several subcategories. In total, as shown in Table 2, the debate commentary taxonomy has 11 fine-grained labels (KS, SE, FC, PO, MR, CQ, and the 5 CPO subtypes), which offer a nuanced view of how commentators respond. Table 2 in Appendix provides frequency statistics of these labels per debate event, confirming that summaries and explanations are most common, but all categories are represented.

The FOMC commentary uses a simpler set of 5 categories reflecting its financial focus. We define labels for: Summary of the Fed's statements; Open Question (similar to CQ, when analysts pose a question or uncertainty); and three sentiment-based Opinion labels – Positive, Neutral, Negative – indicating the tone of the commentator's view on the policy or economic outlook. These sentiment opinions replace the more fine-grained CPO subtypes used in debates, since FOMC commentary often centers on evaluative tone (e.g. optimistic vs pessimistic take on the Fed's message). The earnings call commentary required an even more fine-grained scheme of 10 categories. We include labels for various comparative or contextual analyses that financial journalists provide, such as: comparison with previous company reports, discussion of supply chain details, references to prior quarterly calls, noting market expectations vs actual results, and mentions of competitors' performance. These capture the rich analytical moves typical in earnings analysis. Additionally, earnings commentary labels cover summary of the results, open questions (e.g. uncertainties about guidance), general commentary (uncategorized observations), and sentiment opinions (positive/neutral/negative) about the earnings news. By designing domain-specific labels, we account for differences in commentary style: e.g. debate commentary includes fact-checking political claims, while earnings call commentary often involves comparing numbers to expectations or past quarters.

For the Reddit debate comments, we use a simplified 4-category scheme focusing on how the comment relates to the debate utterance. The labels (drawn from prior work on intertextual links in discussions) are: Agreement, Disagreement, Elab-

oration, or Paraphrase. These indicate whether the Reddit user is agreeing with a candidate's point, disputing it, adding more information or opinion, or simply rephrasing it (often humorously or sarcastically). While not as fine-grained as expert labels, these categories let us study the contrast between expert commentary (which may lean towards factual and analytical responses) and public commentary (which may show more partisanship or humor).

3 Task Design and Evaluation

We consider two primary tasks with our dataset, reflecting the pipeline of a commentary system:

3.1 Commentary Planning

Given a segment of the transcript (e.g. a few sentences of a debate or a turn from the Fed Chair), the model must predict which commentary category an expert would choose for a comment on that segment. This is a multi-class classification task over the label set of the respective domain (e.g. 11-way classification for debates). We evaluate planning performance using standard classification metrics, chiefly accuracy and F1-score. Since the class distribution is imbalanced (certain labels like Key Summary occur more frequently, while others like Commentator's Question are rarer), we report both macro-averaged F1 and micro-F1. The latter emphasizes overall correctness, while macro-F1 highlights performance on less common categories. In our pilot experiments, this task proved very challenging: even powerful LLMs achieved only about 46–49% micro-F1 on debate commentary planning. For example, GPT-4 and Claude 3.5 Sonnet models hovered around 0.5 F1. This indicates that identifying what type of comment to make – essentially, the expert's decision-making – requires deeper understanding of context and likely external knowledge. We expect specialized models or additional context (such as preceding dialogue or world knowledge) to be needed to improve on this task.

3.2 Commentary Generation

Here the goal is to generate the content of a commentary given a transcript segment and a specified commentary label. This reflects producing a particular style of comment (e.g. a fact-check) appropriate to what was said. We treat this as a conditional text generation task. Evaluation of generated commentary is nuanced: we compute automatic metrics like ROUGE (measuring n-gram overlap with the reference expert commentary) and BERTScore (measuring semantic similarity to the reference) to get a quantitative sense of fidelity. However, because commentary is an open-ended task (the model could comment in various valid ways that differ from the single reference), these overlap-based scores tend to be low. Indeed, our pilot tests found ROUGE-1/2 scores in the 0.10 range for even the best LLMs, which underscores that divergent but valid outputs are penalized by reference metrics. We therefore place greater emphasis on human evaluation for generation. We propose to have experts or crowd annotators judge generated commentaries along key dimensions of quality: (a) Importance: does the commentary focus on important or relevant aspects of the segment (as an expert would) rather than trivial details? (b) Expectedness/Novelty: does the commentary provide insight beyond merely restating the transcript (since a good comment should add context or analysis, not just the obvious)? (c) Clarity: is the commentary clearly written and easy to understand? (d) Accuracy: are any factual claims in the commentary correct (this is crucial for fact-checking or explanatory comments). We will use rating scales for these dimensions and also collect an overall preference between different model outputs. Additionally, we plan to utilize LLM-based evaluators for automatic judgment: for example, prompting a strong model to assess a generated commentary for coherence and correctness (drawing on the "news value" criteria from journalism studies). This approach of using LLMs as judges, alongside human evaluation, can help scale the assessment of open-ended generation.

4 Expected Challenges

As a benchmark, we evaluated several cutting-edge LLMs on our tasks. For commentary planning, all models struggled; for instance, Claude's F1 was 0.48, similar to GPT-4, while DeepSeek (a 70B-level open model) was slightly lower, indicating that without fine-tuning, these models often misidentify which strategy to use (e.g. they might summarize when a fact-check was needed, or vice versa). For commentary generation, we experimented with prompting LLMs to produce commentary given segments and target labels. Qualitatively, the models can produce fluent and relevant comments, but often lack the expert precision: e.g. a

fact-check generated by GPT-4 might not actually verify the claim with evidence, or a supposed "market reaction" comment by Claude might be generic since the model doesn't have real financial data. The ROUGE scores around 0.1 for all models reflect that the models' outputs often did not overlap with the reference wording, even if they were topically relevant. This is in stark contrast to, say, news summarization tasks where state-of-the-art models can achieve much higher ROUGE by producing similar summaries. The low scores reaffirm that commentary generation is fundamentally different from summarization: it is a more open-ended, many possible answers problem (especially for opinion and explanation categories). Therefore, we caution against relying solely on reference-based metrics. Instead, our evaluation protocol will use a combination of automatic and human measures, as described, to get a well-rounded picture of performance.

Another challenge is the need for external knowledge. In our dataset, commentators frequently bring in outside information, e.g. citing economic data during a debate or recalling previous statements by the Fed, which a model without retrieval may not know. To encourage research on this, we distinguish between closed-book and open-book commentary generation. A closed-book model must rely only on its internal knowledge and the transcript input, while an open-book model can call a retrieval system or database (for example, retrieve relevant fact-checks or Wikipedia content). We will evaluate both settings. We expect that retrieval-augmented approaches will produce more factual and informative commentary, especially for fact-checking and supplementary explanation categories, at the cost of more complex systems. This setup mirrors real journalists, who often quickly search for data or past news while commenting live.

Overall, our evaluation methodology is designed to capture the multi-dimensional goals of live commentary: factual accuracy, relevance, insight, and timeliness. By providing both the planning labels and the generation task, our dataset allows researchers to decompose the problem.

5 Related Work

Generating live commentary has been explored in limited domains such as sports and games. For example, Ishigaki et al. (2021) generated commentary for racing video games using multimodal inputs, and Marrese-Taylor et al. (2022) proposed open-domain video commentary generation from gameplay. These systems focused on describing visual events, whereas our work deals with discursive events (speeches, discussions) and requires integrating factual knowledge and argumentative context. In the news domain, others have studied generating reader comments or transforming content: e.g., Yang et al. (2019) generated news article comments, and Liu et al. (2024) created *SciNews* to turn scientific papers into lay summaries. Our dataset enables similar grounded generation but in real-time political and financial contexts, which pose unique time-sensitivity and accuracy challenges.

U.S. presidential debates are a rich resource for argument mining and claim analysis. Prior datasets have tackled check-worthy claim detection in debates. The CLEF-2018 CheckThat! lab (Atanasova et al., 2018) introduced tasks to identify which debate statements merit fact-checking. Similarly, *ClaimRank* (Jaradat et al., 2018) prioritized factual claims in debates for fact-checkers. These datasets typically provide binary or priority labels on debate sentences indicating "worth fact-checking." For instance, the Check-Worthy corpus by Patwari et al. (2017) annotated debate sentences with whether they should be checked. However, these resources focus narrowly on factual claims, whereas live commentary covers a broader range of reactions (summaries, opinions, etc.) in real time. Our dataset indeed includes fact-checking commentary labels, but situates them among many other commentary types, providing a more comprehensive view of how debates are analyzed on the fly.

Other work has examined the argumentative structure of debates. The *M-Arg* dataset (Mestre et al., 2021) annotated the 2020 U.S. presidential debates for argument relations (support, attack, neutral) using both text and audio. Goffredo et al. (2023) proposed an argument-based classification of debate content, inspiring parts of our label taxonomy. These efforts treat debates as standalone dialogues to parse or classify, in contrast to our approach of linking debates with external commentary. The CMU *Multivocal* dataset (Jo et al., 2020) integrated social media reactions by categorizing Reddit debate comments into four proposition types. That work illustrated the value of combining debates with crowd commentary, but did not include expert analysis. Our dataset bridges

that gap by including both expert journalist commentary and public Reddit comments for the same debates, enabling direct comparison of institutional versus grassroots discourse. In summary, existing debate datasets each target a slice of the problem (claims, arguments, or crowd opinions), while our dataset provides aligned expert commentary covering fact-checks, summaries, opinions, and more, over multiple election cycles.

Beyond debates, our work draws on NLP research into financial and policy communications. FOMC press conferences (the U.S. Federal Reserve's Q&A sessions after policy meetings) have been studied for their economic impact and rhetoric. For example, prior work analyzed the language of Fed statements to predict market reactions or assess sentiment (e.g., (Zirn et al., 2015; Rohlfs et al., 2016)). Corporate earnings calls are another important domain, with NLP applied to tasks like summarization of call transcripts, extraction of forward-looking statements, and stock movement prediction. Keith and Stent (2019) investigated summarizing earnings calls, and more recent studies (Mukherjee et al., 2022; Huang et al., 2024) use transformer models to analyze financial transcripts. However, these financial NLP works typically operate on monologues or Q&A content alone. Our dataset is novel in that it pairs these financial event transcripts with real-time expert commentary (e.g., from Bloomberg analysts) that interprets and reacts to the content. To our knowledge, this is the first resource to capture how financial experts comment during an unfolding event (press conference or earnings call), adding a layer of analysis akin to real-time summarization plus evaluation.

6 Conclusion

We have presented a new multi-domain dataset for Live Commentary Planning and Generation, covering real-time expert and public commentary on debates, policy press conferences, and earnings calls. This dataset is the first to align transcripts of high-stakes events with time-synchronized expert analyses, annotated with a rich taxonomy of commentary types. By framing both a planning task (deciding what commentary action to take) and a generation task (producing the commentary text), we move toward building systems that not only summarize or classify, but emulate expert commentators in both decision-making and writing. The novelty of our dataset lies in its comprehensive scope and fine granularity: it bridges previously disparate research areas (argument mining, fact-checking, summarization, and discourse analysis) in a unified benchmark. We believe this resource will be highly useful for developing and evaluating the next generation of intelligent assistants capable of providing live analysis. Furthermore, the dataset is extensible: the framework could be applied to other languages (e.g. live translation commentary) or other event types (parliamentary debates, live sports commentary with expert analysts, etc.), enabling cross-cultural and cross-domain studies of real-time commentary.

Looking ahead, we anticipate this dataset will inspire research into planning-enhanced text generation, better integration of external knowledge for live tasks, and evaluation techniques for creative generation. It also offers opportunities for interdisciplinary collaboration with journalism and communication studies, examining how AI can augment or mimic professional commentators. In the era of powerful LLMs, our work highlights that expertise and strategy in generation remain non-trivial to achieve. By providing a challenging benchmark and initial baselines, we set the stage for future innovations in real-time, context-aware text generation. We invite the community to use and build upon our dataset.

References

Pavlin Atanasova, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Stoyan Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. *arXiv:1808.05542*.

Pierpaolo Goffredo, Michele Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112. Association for Computational Linguistics.

Yanlong Huang, Wenxin Tai, Fan Zhou, Qiang Gao, Ting Zhong, and Kunpeng Zhang. 2024. Extracting key insights from earnings call transcripts via information-theoretic contrastive learning. *Information Processing & Management*. To appear.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data.

In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Youngwoo Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020. Machine-aided annotation for fine-grained proposition types in argumentation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1008–1018, Marseille, France. European Language Resources Association.

Katherine A. Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Dan Liu, Yuxi Wang, Jennifer Loy, and Vera Demberg. 2024. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italy. European Language Resources Association and International Committee on Computational Linguistics.

Edison Marrese-Taylor, Yumi Hamazono, Tatsuya Ishigaki, Goran Topić, Yusuke Miyao, Ichiro Kobayashi, and Hiroya Takamura. 2022. Open-domain video commentary generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7326–7339, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ricardo Mestre, Roko Milicic, Stuart E. Middleton, Mark Ryan, Jiechi Zhu, and Timothy J. Norman. 2021. M-Arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhijnan Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM)*, pages 2259–2262.

Christopher Rohlfs, Sunandan Chakraborty, and Lakshminarayanan Subramanian. 2016. The effects of the content of FOMC communications on US treasury rates. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2096–2102, Austin, Texas. Association for Computational Linguistics.

Jacky Visser, Rory Duthie, John Lawrence, and Chris Reed. 2018. Intertextual correspondence for integrating corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: A deep architecture for automatic news comment generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5077–5089, Hong Kong, China. Association for Computational Linguistics.

Cäcilia Zirn, Robert Meusel, and Heiner Stuckenschmidt. 2015. Lost in discussion? tracking opinion groups in complex political discussions by the example of the FOMC meeting transcriptions. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2015)*, pages 747–753, Hissar, Bulgaria.

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

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

Abstract

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

1 Introduction

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

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

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

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

2 Related Work

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

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

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

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

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

3 System Recap

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

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

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

4 Evaluation Setup

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

Dataset	No-Omissions
FA	5.42
CFA	5.21
FI	5.35

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

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

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

5 Results

5.1 Per-dataset No-Omissions

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

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

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

5.2 Pooled Results (D2T-1)

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

6 Discussion

The results highlight a clear profile. DCU-ADAPT-modPB excels in producing grammatical and fluent text with few hallucinations, as reflected in its superior scores on No-Additions, Grammaticality, and Fluency. However, its conservative design results

in lower No-Omissions, especially in CFA, where the system struggled to cover perturbed inputs.

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

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

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

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

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

7 Conclusion

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

Limitations

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

Ethics Statement

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

References

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

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

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

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

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th annual meeting of the Association for Computational Linguistics, pages 311–318.

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

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

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

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

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

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

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

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

Abstract

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

1 System Summary

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

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

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

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

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

2 Official Human Evaluation Results

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

Proceedings of the 18th International Natural Language Generation Conference: Generation Challenges, pages 48–50
October 29–November 2, 2025.

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

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

2.1 Results and Discussion

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

2.1.1 Faithfulness in Factual and Counterfactual Scenarios

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

2.1.2 Grammaticality and Fluency

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

2.2 Comparative Analysis

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

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

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

3 Discussion: a "Reliability Gap"

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

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

4 Conclusion

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

Acknowledgments

This research was funded by DFG grant 389792660 as part of TRR 248 – CPEC.[1] We sincerely thank the GEM'24 shared task organizers for sharing the human evaluation results that helped us showcase the effectiveness of our system.

References

Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021. On training instance selection for few-shot neural text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 8–13, Online. Association for Computational Linguistics.

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

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

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

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

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

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

Author Index