Ask Me Like I'm Human: LLM-based Evaluation with For-Human Instructions Correlates Better with Human Evaluations than Human Judges

Rudali Huidrom and **Anya Belz**
ADAPT Research Centre
Dublin City University
Ireland
{rudali.huidrom,anya.belz}@adaptcentre.ie

Abstract

Human evaluation in NLP has high cost and expertise requirements, and instruction-tuned LLMs are increasingly seen as a viable alternative. Reported correlations with human judgements vary across evaluation contexts and prompt types, and it is hard currently to predict if an LLM-as-judge metric will work equally well for new evaluation contexts and prompts, unless human evaluations are also carried out for comparison. Addressing two main factors contributing to this uncertainty, model suitability and prompt engineering, in the work reported in this focused contribution, we test four LLMs and different ways of combining them, in conjunction with a standard approach to prompt formulation, namely using written-for-human instructions verbatim. We meta-evaluate performance against human evaluations on two data-to-text tasks, and eight evaluation measures, also comparing against more conventional LLM prompt formulations. We find that the best LLM (combination)s are excellent predictors of mean human judgements, and are particularly good at content-related evaluation (in contrast to form-related criteria such as Fluency). Moreover, the best LLMs correlate far more strongly with human evaluations than individual human judges across all scenarios.

1 Introduction

Human evaluation remains the most reliable method for system evaluation in NLP (van Miltenburg et al., 2023b), but its high cost, required expertise, and methodological inconsistencies limit its scalability and reliability (Thomson et al., 2024). The emergence of large language models (Touvron et al., 2023; Chaplot, 2023; Cohere, 2024; Yang et al., 2025) has caused a paradigm shift in text generation and understanding across many domains (Ouyang et al., 2022; Kojima et al., 2022). LLMs are exhibiting state-of-the-art performance in problem-solving and reasoning tasks (Mizrahi et al., 2024; Zhang et al., 2024b). LLMs also hold out the appealing vision of cheaper human-like evaluation, demonstrating adaptability and generalisation capabilities (Li et al., 2024). While individual human judges are subject to inter-rater variability and require multiple annotators for reliability, LLMs may provide more consistent judgements when resources are constrained. 'LLM-as-Judge' approaches do address some of the issues with human evaluation, such as cost and evaluator inconsistency, but their reliability when applied to new tasks needs to be demonstrated via correlation tests with human judgements. In the experiments presented in this paper, we investigate the alignment between human and LLM judgements across a range of criteria for two NLP data-to-text tasks. To standardise prompt formulation, we use the same instructions as those provided in human evaluations, and compare them with more conventional LLM prompts, in conjunction with single models and model combinations of both varying and comparable sizes.

2 Related work

LLM-as-judge has been shown to be an effective approach for assessing a wide range of individual tasks (Liusie et al., 2024). Like other automatic evaluation methods, LLM-as-judge approaches are typically meta-evaluated against human judgement scores, and increasingly on emerging benchmarks, such as HumEval (Chen et al., 2021), SummEval (Fabbri et al., 2021), and MQM (Freitag et al., 2021), used in conjunction with specific evaluation frameworks (Fu et al., 2023; Liu et al., 2023; Liusie et al., 2024, inter alia), or simply with prompts and instructions tailored to the task (Zhang et al., 2024a; Jain et al., 2023; Lin and Chen, 2023; Murugadoss et al., 2025).

In contrast to previous work, we conduct our LLM-as-judge experiments using verbatim human evaluation instructions as a way of standardising prompt formulation. Furthermore, we investigate

LLM-as-judge performance in this setting, comparing with more standard LLM prompt formulations, in meta-evaluation against human judgements on data-to-text tasks.

3 Datasets and Quality Criteria

3.1 WebNLG 2020

WebNLG 2020 is a data-to-text dataset that aligns sets of RDF triples (subject, predicate, object) with text. The English dataset has 1,779 input triple sets in the test set. For the human evaluation, 10% of the test dataset (178 items) was sampled and evaluated on outputs from each team's primary submission (14 submission systems + 3 baseline systems). We use the verbatim criteria from Castro Ferreira *et al.* (2020) which were rated on a scale of 0–100:

Data Coverage: Does the output text include descriptions of all predicates presented in the data?

Relevance: Does the output text describe only such predicates (with related subjects and objects), which are found in the data?

Correctness: When describing predicates which are found in the data, does the text mention correct the objects and adequately introduces the subject for this specific predicate?

Text Structure: Is the text grammatical, well-structured, written in acceptable English?

Fluency: Is it possible to say that the text progresses naturally, forms a coherent whole and it is easy to understand the text?

3.2 ROTOWIRE

ROTOWIRE (Wiseman et al., 2017) is a widely used data-to-text benchmark which contains NBA basketball game statistics and textual summaries for them ($\sim$ 5k instances). The ReproNLP 2023 shared task (Belz and Thomson, 2023) carried out two reproductions (Arvan and Parde, 2023; van Miltenburg et al., 2023a) of the human evaluation in Puduppully and Lapata (2021) which uses this dataset. In the human evaluation, five systems were evaluated on 200 instances per criterion. There are three ratings per item and the participants rank the summaries as either an 'A' or a 'B'. Here too we use the original definitions of the three criteria:

Grammaticality: Is the summary written in well-formed English?

Coherence: Is the summary well structured and well organized and does it have a natural ordering of the facts?

Repetition: Does the summary avoid unnecessary repetition including whole sentences, facts or phrases?

4 LLM-as-Judge Meta-evaluations

4.1 WebNLG'20 LLM-as-judge experiments

In the original WebNLG 2020 evaluation, each paired RDF triple set and system output was evaluated by three human evaluators. We obtain individual scores with each of the following three LLMs, then compute the mean of the three scores from different model and prompt combinations:

- J_H: LLM judgements using as the prompt the verbatim instructions from the original human evaluation in WebNLG 2020.
- J_{C+D}: LLM judgements using as the prompt conventional minimal zero-shot LLM prompts also incorporating the verbatim evaluation criterion definitions.
- J_{C-D}: Same as J_{C+D} minus the definitions.
- H: For comparison, we also test single human judgements from WebNLG'20 as predictors.

We use the following models (details Appendix C):

- Llama3-8B-Instruct (Touvron et al., 2023)
- Mistral-7B-Instruct-v0.2 (Chaplot, 2023)
- C4AI Command R+ (Cohere, 2024)

4.2 Rotowire LLM-as-judge experiments

In the original ROTOWIRE evaluation, system summaries were evaluated by three human evaluators. We obtain individual ratings with each of our three LLMs, then compute the majority vote of the three ratings. In this context, we use just the for-human instructions as in the original human evaluation. We test the correlations between the following LLM (combination)s and human judgements:

- H_1 and H_2: Two sets of human judgements obtained from two reproductions of Puduppully and Lapata (2021).
- J_{H_V}: Majority vote of LLM judgements by models of varying sizes (7B, 8B, 104B) and using the same human instructions (same models as in the WebNLG 2020 tests).
- J_{H_C}: Majority vote of LLM judgements on models of comparable sizes (two 7Bs and one 8B) and using the same human instructions. These are Llama3-8B-Instruct, Mistral-7B-Instruct-v0.2 and Qwen2.5-7B-Instruct-1M.

We use the same models as for WebNLG in the J_{H_V} tests, and replace the Cohere model with Qwen2.5-7B-Instruct-1M. (Yang et al., 2025)

	Correctness				Data Coverage				Fluency				Relevance				Text Structure			
	H	J_H	J_{C+D}	J_{C-D}	H	J_H	J_{C+D}	J_{C-D}	H	J_H	J_{C+D}	J_{C-D}	H	J_H	J_{C+D}	J_{C-D}	H	J_H	J_{C+D}	J_{C-D}
AAI	93.53	97.62	97.04	95.16	94.39	96.29	97.37	91.21	90.29	95.58	94.59	90.67	95.20	99.57	97.69	92.89	92.95	97.19	95.40	88.47
F17	90.14	97.13	97.88	94.54	92.07	94.67	97.72	90.56	80.94	95.11	94.70	90.29	92.59	99.62	98.25	92.15	85.74	97.14	95.41	87.45
F20	92.31	97.78	97.99	95.47	93.42	96.32	97.96	91.49	82.6	95.76	95.09	91.19	94.31	99.97	98.28	93.19	87.89	97.31	95.58	88.53
bt5	93.58	96.57	95.71	94.33	93.84	95.54	96.23	90.69	88.69	94.66	93.75	90.23	95.22	99.68	97.29	92.26	91.91	97.29	95.04	87.73
cuni	91.59	95.63	95.30	95.06	93.29	94.71	96.19	91.51	87.64	94.18	92.94	90.84	94.56	99.67	96.92	93.03	90.75	97.2	94.42	88.48
CGT	89.85	94.56	96.17	94.35	91.23	93.86	97.02	90.63	84.82	92.83	93.21	90.07	93.37	99.40	98.27	92.28	87.88	96.98	94.91	87.48
D-SGU	92.49	96.12	95.44	93.66	95.32	95.08	96.27	90.05	78.59	93.31	90.92	88.61	94.86	99.8	97.28	91.46	83.50	96.52	92.38	86.33
FB-AI	92.70	97.35	97.31	95.18	93.17	96.30	97.50	91.25	90.84	95.87	94.98	90.86	93.9	99.88	98.06	92.90	93.09	97.51	95.67	88.48
H_Lab	80.76	85.82	88.54	90.48	84.74	86.93	92.52	88.24	75.21	82.78	83.92	85.24	85.27	96.11	94.55	88.59	80.22	92.16	88.16	82.86
NILC	76.70	77.64	81.75	88.34	81.61	79.28	86.64	84.58	74.85	77.17	78.93	82.82	83.52	91.87	90.56	84.74	80.46	88.62	86.88	80.98
NUIG	92.05	96.06	95.49	95.02	92.06	95.18	96.53	91.41	88.90	94.68	93.83	90.61	94.06	99.14	97.31	92.85	91.59	97.35	95.06	88.23
O-NLG	74.98	74.29	77.35	85.00	79.96	77.68	82.68	83.94	75.68	73.12	74.83	79.90	79.89	88.03	86.74	81.65	80.46	85.14	84.43	78.71
OSU	93.41	96.57	95.78	95.16	95.12	95.48	96.67	91.14	90.07	95.50	93.83	90.72	94.62	99.31	97.38	93.04	92.44	97.41	95.10	88.65
RALI	92.13	97.54	96.52	94.56	95.20	96.20	96.69	90.82	77.76	94.86	92.53	89.83	94.81	99.74	97.52	92.48	81.84	97.07	94.12	87.46
TGEN	88.63	95.64	95.02	96.29	88.18	94.62	95.55	92.64	86.16	94.43	92.83	91.28	92.64	99.46	96.99	94.14	89.04	97.31	94.42	89.01
UPC	74.37	79.59	83.86	89.27	75.85	81.59	89.06	87.61	72.28	77.63	79.57	84.00	82.05	94.66	93.68	87.68	78.50	88.82	86.46	81.77
W-REF	94.15	97.59	97.64	95.01	95.44	95.99	97.71	91.02	89.85	95.54	95.38	90.67	94.39	99.80	98.35	92.96	92.11	97.28	95.83	88.16
Avg	88.43	92.56	93.22	**93.35**	90.29	92.10	**94.72**	89.93	83.25	**90.77**	90.34	88.7	91.49	**97.98**	96.18	91.08	87.08	**95.19**	92.90	86.4

Table 1: System-level average scores for each quality criterion by WebNLG'20 human judges (H), average over Llama3-8B-Instruct/Mistral-7B-Instruct-v0.2/Command R+ prompted with full human instructions (J_H); conventional zero-shot prompt with (J_{C+D}) and without definitions (J_{C-D}). System names (rows) with length > 4 letters are shortened by concatenating the first letter or first two letters with the last two/three letters.

	Single Human Judges Avg	Human Instructions as prompt mean of 3 scores by:				Zero-shot + original definitions mean of 3 scores by:				Zero-shot - original definitions mean of 3 scores by:			
		Mistral	Llama	CRplus	Mistral+ Llama+ CRplus	Mistral	Llama	CRplus	Mistral+ Llama+ CRplus	Mistral	Llama	CRplus	Mistral+ Llama+ CRplus
Correctness	0.69	0.93	0.94	0.99	**0.97**	0.72	0.93	0.98	0.95	0.90	0.25	0.98	0.92
Data Coverage	0.68	0.89	0.86	**0.96**	0.93	0.62	0.84	**0.96**	0.88	0.77	0.21	0.93	0.79
Fluency	0.68	0.67	0.75	**0.81**	0.78	0.48	0.84	**0.81**	0.80	0.74	0.68	0.79	0.79
Relevance	0.69	0.85	0.90	**0.98**	0.94	0.67	0.93	0.96	0.91	0.93	0.66	0.96	0.93
Text Structure	0.69	0.49	0.70	0.79	0.76	0.16	0.79	**0.87**	0.83	0.79	0.74	0.79	0.82

Table 2: Pearson's correlations with the aggregated WebNLG'20 human scores, achieved by single human judges and different LLMs.

4.3 Common details

We execute the above prompts as zero-shot inference prompts on the above LLMs. Moreover, we run the experiments with three different seeds (42; 1738; 1,234), meaning each score in tables below is the average of the outputs from the different seed runs. All experiments use English data.

5 Results and Analysis

5.1 Mean scores

Table 1 presents the system-level average scores per evaluation criterion for WebNLG. We observe that human evaluators and LLM judges generally agree with each other, with AAI, F17, F20, OSU, and W-REF often emerging as top performers and, O-NLG and UPC consistently rated lower by both human and LLM judges across multiple criteria.

Moreover, the averages of system-level scores (last row) by LLMs are higher than those by humans in all cases except three averages produced by the zero-shot prompt without definitions (J_{C-D}).

Table 3 presents the system-level average scores per evaluation criterion for the two Rotowire human evaluations (H_1, H_2), and the two types of majority vote, one with a much larger model in the mix (J_{H_V}), and one with similar sized models (J_{H_C}). Human and LLM judges agree on the high performance of the Gold system, although H_1 uniquely favours the Template system. Additionally, while J_{H_V} and J_{H_C} yield similar evaluations for top-performing systems, J_{H_C} tends to assign slightly higher scores for lower-performing systems (e.g., Template) in Coherence and Repetition.

5.2 Correlations with human judgements

Table 2 reports the correlations with the original WebNLG'20 human judgements achieved by: (i) individual human judges on average, (ii) each of the LLM model (combination)s. Strikingly, individual human judges have far lower agreement with the mean of the other judges (on the same outputs) than the LLMs. Another clear result is that the different models are affected very differently by

	Coherence				Repetition				Grammaticality			
	H_1	H_2	J_{H_V}	J_{H_C}	H_1	H_2	J_{H_V}	J_{H_C}	H_1	H_2	J_{H_V}	J_{H_C}
Gold	49.79	**56.25**	**70.00**	**70.00**	49.16	**52.92**	**70.83**	**73.75**	54.62	**57.08**	**70.83**	**64.58**
Template	**62.76**	40.00	18.75	24.58	**72.15**	47.08	22.92	26.25	**58.58**	38.33	32.08	42.08
ED+CC	42.50	46.25	42.08	41.67	36.97	47.50	44.17	41.67	40.17	45.83	37.50	40.42
Hier	44.77	54.58	60.42	56.67	42.62	50.42	56.25	51.67	45.19	54.58	52.92	49.17
Macro	50.21	52.92	58.75	57.08	49.15	52.08	55.83	56.67	51.48	54.17	56.67	53.75

Table 3: System-level average scores for each quality criterion by two sets of Rotowire human judges (H_1, H_2), average majority vote by varying-size models Llama3-8B-Instruct/Mistral-7B-Instruct-v0.2/Command R+ (J_{H_V}), and average majority vote by Llama3-8B-Instruct/Mistral-7B-Instruct-v0.2/Qwen2.5-7B-Instruct-1M (J_{H_C}).

	H_1	H_2	J_{H_V}	J_{H_C}
Coherence				
H_1	1.000	-0.585	-0.626	-0.548
H_2	-0.585	1.000	0.992	0.982
J_{H_V}	-0.626	0.992	1.000	0.993
J_{H_C}	-0.548	0.982	0.993	1.000
Grammaticality				
H_1	1.000	-0.185	0.134	0.358
H_2	-0.185	1.000	0.931	0.814
J_{H_V}	0.134	0.931	1.000	0.969
J_{H_C}	0.358	0.814	0.969	1.000
Repetition				
H_1	1.000	-0.279	-0.620	-0.482
H_2	-0.279	1.000	0.899	0.936
J_{H_V}	-0.620	0.899	1.000	0.981
J_{H_C}	-0.482	0.936	0.981	1.000

Table 4: Pearson's correlation matrix for Rotowire / Coherence, Grammaticality & Repetition.

differences in prompts: all perform broadly similarly with the verbatim human instructions; Mistral scores collapse when human instructions are removed and definitions are retained, but recover when the definitions are also removed; and Llama scores are unaffected by the removal of human instructions, but collapse when the definitions are also removed. The Command R+ models does best with the human instructions, but largely retains its performance under the other two conditions.

Table 4 shows the complete correlation matrices between the two sets of Rotowire human judges and the two majority-voting combinations of LLMs, for each of the three evaluation criteria. Here, the most striking result is the stark discrepancy between the two sets of human judges: H_1 has a medium strong *negative* correlation with both H_2 and the LLMs for Coherence, weak or no correlation for Grammaticality, and weak or medium *negative* correlation for Repetition. In contrast H_2 and LLM combinations all agree strongly with each other. H_1 and H_2 also produced different reproducibility assessments compared to the original evaluation by Puduppully and Lapata (2021), as reported in the ReproNLP 2023 shared task report (Belz and Thomson, 2023).

In this situation, where one set of human evaluations disagrees with another, we have no basis for deciding which of the two gives a truer picture: either H_2 is right or H_1 is right, but they can't both be right. In this situation, a new role emerges for LLMs: as sanity checkers when human evaluations disagree. We discuss this further in the next section, and in a forthcoming paper (Huidrom and Belz, 2025).

6 Discussion and Conclusion

We have presented results for experiments with LLM-as-judge approaches for two types of data-to-text tasks and eight evaluation methods, using as a way of standardised prompt formulations the verbatim human instructions from previous evaluations. These were shown to work better than more conventional prompt formulations in all scenarios, irrespective of task or the length of input/output.

An unexpected discovery was that LLMs can serve as sanity checkers for human evaluations. The ReproNLP shared task organisers had no basis for deciding which of two reproductions of Puduppully and Lapata (2021) they reported was right: either Repro 1 (H_2 in this paper) was right and the work had excellent reproducibility, or Repro 2 (H_1) was right and it had terrible reproducibility. Because both of our LLM majority votes strongly agreed with Repro 1 and strongly disagreed with Repro 2, the indication is that Repro 1 (H_2) gave the better results out of the two reproductions.

Overall, we have found our best LLMs to be highly reliable predictors of human evaluations, and to benefit from human-type detailed evaluation instructions. The result that individual human judges correlate far less well with overall human judgements than LLMs do, implies that if the choice is between a small number of human judges and an LLM you are better off using the LLM.

Limitations

The experiments conducted showed promising alignment between human and LLM evaluations. Our evaluation covered only a limited set of models and tasks, so our findings are confined to those.

Ethics Statement

As a paper that meta-evaluates existing human evaluation tasks using the same and custom instructions, the risk associated with this study was minimal.

Acknowledgments

We thank all the reviewers for their valuable feedback and advice. Huidrom's work is supported by the Faculty of Engineering and Computing, DCU, via a PhD studentship. Both authors benefit from being members of the SFI Ireland funded ADAPT Research Centre.

References

Mohammad Arvan and Natalie Parde. 2023. Human evaluation reproduction report for data-to-text generation with macro planning. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 89–96, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Simone Balloccu, Anya Belz, Rudali Huidrom, Ehud Reiter, João Sedoc, and Craig Thomson. 2024. Proceedings of the fourth workshop on human evaluation of nlp systems (humeval)@ lrec-coling 2024. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*.

Anya Belz, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina. 2021. Proceedings of the workshop on human evaluation of nlp systems (humeval). In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*.

Anya Belz, Maja Popović, Ehud Reiter, and Anastasia Shimorina. 2022. Proceedings of the 2nd workshop on human evaluation of nlp systems (humeval). In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*.

Anya Belz, Maja Popović, Ehud Reiter, Craig Thomson, and João Sedoc. 2023. Proceedings of the 3rd workshop on human evaluation of nlp systems. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*.

Anya Belz and Craig Thomson. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Cohere. 2024. Introducing command r+: A scalable llm built for business. *https://cohere.com/blog/command-r-plus-microsoft-azure*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Rudali Huidrom and Anya Belz. 2025. Using llm-as-judge evaluation for sanity-checking results and reproducibility of human evaluations of nlp systems. In *Proceedings of the 4th Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, Vienna, Austria. Association for Computational Linguistics.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Adian Liusie, Vatsal Raina, Yassir Fathullah, and Mark Gales. 2024. Efficient llm comparative assessment: a product of experts framework for pairwise comparisons. *arXiv preprint arXiv:2405.05894*.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2025. Evaluating the evaluator: Measuring llms' adherence to task evaluation instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19589–19597.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Craig Thomson, Ehud Reiter, and Belz Anya. 2024. Common flaws in running human evaluation experiments in nlp. *Computational Linguistics*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. 2023a. How reproducible is best-worst scaling for human evaluation? a reproduction of 'data-to-text generation with macro planning'. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 75–88, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. 2023b. How reproducible is best-worst scaling for human evaluation? a reproduction of 'data-to-text generation with macro planning'. *Human Evaluation of NLP Systems*, page 75.

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

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Kaiqi Zhang, Shuai Yuan, and Honghan Zhao. 2024a. Talec: teach your llm to evaluate in specific domain with in-house criteria by criteria division and zero-shot plus few-shot. *arXiv preprint arXiv:2407.10999*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024b. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

A WebNLG 2020 Dataset

The WebNLG+ 2020 Challenge focused on (i) mapping RDF triples to generate English or Russian texts (generation) and (ii) converting English or Russian texts into RDF triples (semantic parsing). Our work addresses the generation task for English. The English WebNLG 2020 dataset (version 3.0) comprises 13,211/1,667/1,779 triple sets in the train, dev, and test splits, respectively, with triple sizes ranging from one to seven and 19 DBpedia categories, three of which are unseen in the training set. The challenge involved 15 teams submitting 48 system runs, with 14 teams focusing on English data and six on Russian data.

For the human evaluation, 10% of the test dataset was sampled (178 samples) and evaluated on each team's primary system submission. (Castro Ferreira et al., 2020) recruited 109 annotators via Ama-

zon Mechanical Turk, providing them with instructions (criteria on a 0–100 slider scale), RDF triples, and system outputs. Each sample received three annotations.

B ROTOWIRE Dataset

The ReproHum initiative (Belz et al., 2021, 2022, 2023; Balloccu et al., 2024) curated two reproductions (Arvan and Parde, 2023; van Miltenburg et al., 2023a), of the human evaluation in Puduppully and Lapata (2021) which uses the ROTOWIRE dataset. Five systems were evaluated over three criteria on 200 instances per criteria. In total, there are 600 instances across all criteria. There are three ratings per item and the participants can only respond using the characters 'A' or 'B' to indicate their preference over the summaries. There were a total of 216 participants in the first reproductions and 262 participants in the second reproductions. The original study does not provide raw human evaluation scores, which is why we used the reproduced scores for comparison in our work.

C Models Used

Below are the models we used in our experiments; they were selected for being open-source, instruction-tuned LLMs with high ratings on Hugging Face.

- Llama3-8B-Instruct:[1] Meta's Llama 3 series model in the smaller 8B parameter size is pretrained, instruction-tuned, but also optimised for dialogue-based applications.
- Mistral-7B-Instruct-v0.2:[2] Mistral-7B-Instruct is a language model designed to follow instructions, generate creative text, and handle requests, fine-tuned from Mistral-7B-v0.2 using a diverse range of public conversation datasets.
- C4AI Command R+:[3] Cohere's open-weights research release of a 104B parameter model; a multilingual model evaluated in 10 languages for performance, and optimised for a variety of tasks including reasoning, summarization, and question answering.
- Qwen2.5-7B-Instruct-1M:[4] Alibaba's Qwen series model in the smaller 7B parameter size is fine-tuned, instruction-tuned and is optimised to handle long-context tasks while maintaining its capability in short tasks.

D Experiment Setup

We briefly outline the experimental setup used in all of our experiments in this section. We use three large language models for our experiments: Meta-Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.2 and c4ai-command-r-plus-4bit. For hyperparameters, we set temperature to 0.001, maximum length to 1024 for WebNLG'20 & 128 for ROTOWIRE and top p to 1. The choice of our hyperparameters is to produce near-deterministic outputs while preserving subtle probabilistic distinctions in the model's token preferences. We quantise the models to 4-bit and use one rtxa6000/a100 GPU for the execution of our experiments. The cumulative GPU time required for our experiments was a little over 150 GPU hours.

E Experimental Grid

For WebNLG 2020: {English}x{Llama3-8B-Instruct, Mistral-7B-Instruct-v0.2, command-r-plus-4bit}x{zero-shot}x{seeds: 42, 1738, 1234}x{Evaluator(s) set-up: one LLM as one evaluator on (a) same instructions as the human evaluation, (b) custom minimal zero-shot prompt with original definitions included, (c) custom minimal zero-shot prompt without original definitions included}.

For ROTOWIRE: {English}x{Llama3-8B-Instruct, Mistral-7B-Instruct-v0.2, command-r-plus-4bit, Qwen2.5-7B-Instruct-1M}x{zero-shot}x{seeds: 42, 1738, 1234}x{Evaluator(s) set-up: one LLM as one evaluator on same instructions as the human evaluation across (a) models of varying sizes, (b) models of comparable sizes}.

F Prompts

We present the prompt used in our experiments in this section. In particular, we outline the general instruction used for all LLMs, we present the prompt template for each LLM. All of this can be found in Tables 5–7.

[1] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

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

[4] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M

G Use of AI Assistants for Writing

We use AI Assistants to sanity check our writing. Grammarly was used for grammar checking, QuillBot (mostly) and ChatGPT (sometimes) were used for rephrasing.

Common Template for All Prompts for J_H	
{task_desc}	Please (i) follow the instructions, (ii) be honest and fair in your judgements, (iii) try to be as correct as possible in your conclusions. For example, the text would generally get a score higher than 0 for Correctness if at least some objects in it are introduced correctly. Similarly, the text would not be rated with 100 for Correctness if at least one object is not introduced correctly.
{task_instr}	Task Instructions: You are given a piece of data and a text that describes data. Below you will find statements that relate to the text. Please rate each of these statements by moving the slider along the scale where 0 stands for 'I do not agree', and 100 stands for 'I fully agree'.
{data}	DATA:
{desc}	DESCRIPTION:
{statement}	How well do you agree with the following statements?
{datacoverage_criteria}	Data Coverage: The text contains all predicates from the data and does not miss any predicates shown in the data.
{relevance_criteria}	Relevance: The text contains only known/relevant predicates, which are found in the data. The text does not contain any unknown/irrelevant/unrecognizable predicates.
{correctness_criteria}	Correctness: When describing information about relevant predicates (those which are in both data and text), the text depicts them with correct/proper objects. Also, the text correctly introduces the subject.
{textstr_criteria}	Text Structure: The text is written in good English, i.e., it is free from grammatical errors and well-structured.
{fluency_criteria	Fluency: The text sounds logically correct and forms a coherent whole. There are no parts of the text you would change to make it sound better. The text forms a nice narrative.
{feedback}	Write your feedback in the field below if you have any (not necessary):
Llama3-8B-Instruct Prompt	
Special tokens	{llama3_bos}: ⟨|begin_of_text|⟩ ; {llama3_eos}: ⟨|end_of_text|⟩ ; {llama3_sot}: {{; {llama3_eot}: }}
Template	{llama3_bos} {llama3_sot}{task_description}{task_instruction}{data}{triples} {description}{verb} {statement}{datacoverage}{relevance}{correctness} {textstructure}{fluency}{feedback} {llama3_eot}{llama3_eos} Data Coverage: Relevance: Correctness: Text Structure: Fluency:
Mistral-7B-Instruct-v0.2 Prompt	
Special tokens	{mistral_bos}: ⟨s⟩ ; {mistral_eos}: ⟨/s⟩ ; {mistral_sot}: [INST] ; {mistral_eot}: [/INST]
Template	{mistral_bos}{mistral_sot} {task_description}{task_instruction}{data}{triples} {description}{verb} {statement}{datacoverage}{relevance}{correctness} {textstructure}{fluency}{feedback}{mistral_eot}{mistral_eos} Data Coverage: Relevance: Correctness: Text Structure: Fluency:
Command-r-plus-4bit Prompt	
Special tokens	{commandrplus_instruction}: ## Instructions\n; {commandrplus_input}: ## Input\n; {commandrplus_output}: ## Output\n; {commandrplus_criterion}: ## Criterion\n
Template	{commandrplus_instruction}{task_description} {task_instruction}{commandrplus_input}{data}{triples} {commandrplus_output}{description}{verb}{commandrplus_criterion} {statement}{datacoverage}{relevance}{correctness} {textstructure}{fluency}{feedback}Output: Data Coverage: Relevance: Correctness: Text Structure: Fluency:

Table 5: Human Evaluation Guidelines from WebNLG 2020 given to the LLMs.

Common Template for All Prompts for J_{C+D} & J_{C-D}	
{our_task_desc}	You are an evaluator. Please read the instructions carefully and provide your judgements honestly and accurately.
{zs_minimal}	Rate the following input triple(s) and text that describes the input triple(s) on a scale from 0 to 100 based on the following criteria:
{input_triples}	Input Triple(s):
{text}	Text:
{datacoverage_criteria}	Data Coverage: The text contains all predicates from the data and does not miss any predicates shown in the data.
{relevance_criteria}	Relevance: The text contains only known/relevant predicates, which are found in the data. The text does not contain any unknown/irrelevant/unrecognizable predicates.
{correctness_criteria}	Correctness: When describing information about relevant predicates (those which are in both data and text), the text depicts them with correct/proper objects. Also, the text correctly introduces the subject.
{textstr_criteria}	Text Structure: The text is written in good English, i.e., it is free from grammatical errors and well-structured.
{fluency_criteria	Fluency: The text sounds logically correct and forms a coherent whole. There are no parts of the text you would change to make it sound better. The text forms a nice narrative.
Llama3-8B-Instruct Prompt	
Special tokens	{llama3_bos}: ⟨|begin_of_text|⟩ ; {llama3_eos}: ⟨|end_of_text|⟩ ; {llama3_sot}: {{; {llama3_eot}: }}
Template	{llama3_bos} {llama3_sot}{our_task_desc}{zs_minimal} {datacoverage}{relevance}{correctness}{textstructure}{fluency} {input_triples}{triples} {text}{verb}{llama3_eot}{llama3_eos} Output: Data Coverage: Relevance: Correctness: Text Structure: Fluency:
Mistral-7B-Instruct-v0.2 Prompt	
Special tokens	{mistral_bos}: ⟨s⟩ ; {mistral_eos}: ⟨/s⟩ ; {mistral_sot}: [INST] ; {mistral_eot}: [/INST]
Template	{mistral_bos}{mistral_sot}{our_task_desc}{zs_minimal} {datacoverage}{relevance}{correctness}{textstructure}{fluency} {input_triples}{triples} {text}{verb}{mistral_eot}{mistral_eos} Output: Data Coverage: Relevance: Correctness: Text Structure: Fluency:
Command-r-plus-4bit Prompt	
Special tokens	{commandrplus_instruction}: ## Instructions\n; {commandrplus_criterion}: ## Criterion\n {commandrplus_input}: ## Input\n; {commandrplus_output}: ## Output\n
Template	{commandrplus_instruction}{our_task_desc}{zs_minimal} {commandrplus_criterion}{datacoverage}{relevance}{correctness}{textstructure}{fluency} {commandrplus_input}{input_triples}{triples} {commandrplus_output}{text}{verb} Output: Data Coverage: Relevance: Correctness: Text Structure: Fluency:

Table 6: Custom zero-shot instructions given to the LLMs.
{datacoverage}{relevance}{correctness}{textstructure}{fluency} is used only for instructions with definitions.

Common Template for All Prompts for J_{H_V} & J_{H_C}	
{summaries}	Summaries
{sys_summaries}	System Summaries
{A}	A:
{B}	B:
{rank_criteria}	Ranking Criteria
{Criteria}	Coherence or Grammaticality or Repetition
{answer}	Answers
{best}	Best:
{worst}	Worst:
{analysis}	Analysis
System-level Prompt	
{gen_instr_rotowire}	You are a native speaker of English or a near-native speaker who can comfortably comprehend summary of NBA basketball games written in English.
{task_head_rotowire}	Evaluate Sports Summaries of (NBA) basketball games.
{task_instr_rotowire}	Your task is to read two short texts which have been produced by different automatic systems. These systems typically take a large table as input which contains statistics of a basketball game and produce a document which summarizes the table in natural langauge (e.g., talks about what happened in the game, who scored, who won and so on). Please read the two summaries carefully and judge how good each is according to the following criterion:
{task_desc_rotowire}	This task contains validation instances (for which answers are known) that will be used for an automatic quality assessment of submissions. Therefore, please read the summaries carefully.
System Prompt:	{gen_instr_rotowire} {task_head_rotowire} {task_instr_rotowire} {task_desc_rotowire}
Llama3-8B-Instruct Prompt	
Special tokens	{llama3_bos}: ⟨|begin_of_text|⟩ ; {llama3_eos}: ⟨|end_of_text|⟩ ; {llama3_sot}: {{; {llama3_eot}: }}
Template	{llama3_bos}{llama3_sot}{summaries}{sys_summaries}{A}{a} {B}{b} {rank_criteria}{Criteria}{answer}{best} {worst} {analysis}{llama3_eot}{llama3_eos}
Mistral-7B-Instruct-v0.2 Prompt	
Special tokens	{mistral_bos}: ⟨s⟩ ; {mistral_eos}: ⟨/s⟩ ; {mistral_sot}: [INST] ; {mistral_eot}: [/INST]
Template	{mistral_bos}{summaries}{sys_summaries}{A}{a} {B}{b} {rank_criteria}{Criteria}{answer}{best} {worst} {analysis}{mistral_eot}{mistral_eos}
Command-r-plus-4bit Prompt	
Special tokens	{commandrplus_instruction}: ## Instructions\n; {commandrplus_criterion}: ## Criterion\n {commandrplus_input}: ## Input\n; {commandrplus_output}: ## Output\n
Template	{commandrplus_instruction}{summaries}{sys_summaries} {commandrplus_input}{A}{a} {B}{b} {commandrplus_criterion}{rank_criteria}{Criteria} {commandrplus_output}{answer}{best} {worst} {analysis} Output: Best: Worst:
Qwen2.5-7B-Instruct-1M Prompt	
Special tokens	-
Template	{summaries}{sys_summaries}{A}{a} {B}{b} {rank_criteria}{Criteria}{answer}{best} {worst} {analysis} Output: Best: Worst:

Table 7: Human Evaluation Guidelines from Puduppully and Lapata (2021) given to the LLMs.