LLM Multi-Agent Systems for Data-to-Text Generation from Large Triple Sets

Chinonso Cynthia Osuji[♡], **Simon Mille**[♡], **Mark Andrade**[♡], **Jane Adkins**[♡], **Ornait O'Connell**[♡], **Elaine Uí Dhonnchadha**[♠], **Bláithín Heffernan**[♡], **Fírinne Nic an tSaoir**[♡], **Anya Belz**[♡], **Thiago Castro Ferreira**[♣], **Brian Davis**[♡]

[♡] ADAPT Research Centre, Dublin City University, Ireland

[♠] Trinity College Dublin, Ireland [♣] Fluminense Federal University, Brazil

firstname.lastname@adaptcentre.ie, thiago.castro.ferreira@gmail.com

Abstract

Generating coherent, semantically accurate text from large structured inputs remains a persistent challenge in data-to-text generation, as single-step LLM mappings from data-to-text limit control over discourse structuring and amplify hallucinations and omissions as input size grows. We introduce a new dataset of extended DBpedia triple sets (up to 199 triples per input), and a modular multi-agent framework: specialised LLM agents handle content ordering, text structuring, and surface realisation under the supervision of an orchestrator and guardrail control loop. The system generates multi-paragraph outputs in English and Irish (low-resource). We compare a three-worker multi-agent configuration against a single-worker multi-task variant and a strong end-to-end baseline. Quality is assessed via human evaluation and LLM-as-a-judge (with truncation-based sanity checks). Results show slightly superior coherence for the multi-agent approach in both languages, with statistically significant inter-rater correlation over all criteria for English and no statistically significant correlation for Irish. Human-LLM alignment is very weak overall, thus exposing key limits in scalable NLG evaluation.

1 Introduction

Modern large language models (LLMs) produce fluent textual descriptions from structured data representations across multiple languages. However, their performance on long, densely structured inputs remains underexplored. Existing data-to-text (D2T) benchmarks that do not involve content selection predominantly feature short input-output pairs designed for sentence or paragraph-level generation, with reference texts spanning only a few dozen tokens (Novikova et al., 2017; Gardent et al., 2017; Parikh et al., 2020). For contemporary LLMs, these datasets present limited challenges in terms of input length, information coverage, and discourse-level complexity.

Meanwhile, advances in long-context modeling have dramatically increased nominal context windows, with many LLMs now supporting inputs of 32k tokens or more (Gao et al., 2025). Despite these architectural improvements, benchmark evaluations reveal significant limitations in effective long-document generation tasks (Li et al., 2024a; Liu et al., 2024b, 2025; Wu et al., 2025b). Performance typically degrades as context length grows, with models underutilizing information from the middle portions of input sequences (Liu et al., 2024a). Furthermore, long-form generation exhibits increased incoherence or hallucination in later sections of the output (Yang et al., 2025; Xu et al., 2023). This indicates a need for enhanced control mechanisms, particularly for document-level D2T tasks requiring sustained factual coverage and discourse coherence across multiple paragraphs. This challenge has spurred considerable interest in LLM-based agents and multi-agent systems (Xi et al., 2025; Li et al., 2024b), in the context of which recent surveys identify role specialization or hierarchical communication as critical components for handling complex, constrained generation tasks (Guo et al., 2024; Gu et al., 2024; Du et al., 2025; Wang et al., 2025).

This paper addresses the lack of long-context datasets and of exploration of multi-agent-based architectures in D2T generation by creating a new dataset of DBpedia triple sets of up to 199 triples, and testing three system configurations: (i) a three-worker multi-agent system with specialized agents for content ordering, text structuring, and surface realization; (ii) a multi-task single-worker variant consolidating all three stages within a single prompt; and (iii) a strong end-to-end baseline. We evaluate outputs in both high resource (English) and low resource (Irish) settings. For intrinsic text quality (e.g. `Fluency`), our evaluation methodology uses both LLM-as-a-judge assessment and expert human ratings. For semantic accuracy crite-

Findings of the Association for Computational Linguistics: ACL 2026, pages 34261–34275

ria (e.g. `No-Omissions`), LLM-as-a-judge is used and LLM sensitivity to different percentages of missing contents in the input and the output is assessed. All code, prompts, data, and results can be found at https://github.com/NonsoCynthia/D2T-Longer-Text-Agent-System.

2 Related Work

Constrained long-form generation has been cast as an iterative cognitive workflow that includes planning, drafting, monitoring, and reviewing (Wan et al., 2025; Wu et al., 2025a) or as heterogeneous recursive planning, where an agent framework interleaves task decomposition and execution across retrieval, reasoning, and composition (Xiong et al., 2025). Multi-agent collaboration extends these ideas to document-level rewriting and simplification; related work on ordered cooperation patterns, where agents communicate in sequence and downstream agents attend only to upstream outputs, indicates that such structured turn taking improves efficiency and helps maintain more organised multi-agent discussions (Xi et al., 2025). However, recent agent-based approaches to D2T generation are evaluated on relatively short inputs and do not target document-level, multilingual long-context generation (Osuji et al., 2025b; Lango and Dušek, 2025).

On the data side, existing foundational D2T datasets exhibit relatively short average output lengths: E2E (22.67 tokens) (Novikova et al., 2017), WebNLG (22.69 tokens) (Castro Ferreira et al., 2020), ToTTo (17.4 tokens) (Parikh et al., 2020), WeatherGOV (28.70 tokens) (Liang et al., 2009), or WikiBio (26.1 tokens) (Lebret et al., 2016). While datasets like RotoWire (337.10 tokens) (Wiseman et al., 2017) offer longer outputs, they typically necessitate additional content selection components, which complicates application to direct D2T generation. In recent work (Osuji et al., 2025b), we proposed a dataset of comparable output size to RotoWire (≈300 tokens), with up to 69 triples in the input; in this paper, we used input triple sets of up to 199 triples.

3 Methodology

We model the D2T generation task over sets of RDF triples that represent long structured inputs as follows. Each input is a DBpedia triple set $X = \{(s_i, p_i, o_i)\}_{i=1}^{N}$, where s_i, p_i, and o_i denote the Subject, Property, and Object of the i-th triple, and all triples are represented in English (see Figure 3 in Appendix A for a sample triple set). Given X and a target language $\ell \in \{\text{en}, \text{ga}\}$, the system must produce a document-level text Y_ℓ in the chosen language that is semantically complete with respect to X (and only X), and of high discourse quality.

The underlying LLMs are used in a frozen setting, so we do not optimize their weights directly. Instead, we control their behaviour through prompt design, role-specific instructions, and the modular multi-agent architecture in Algorithm 1 in App. B. For all experiments, we set the decoding temperature to 0 to encourage deterministic behaviour.

Both for generating texts and automatically assessing their quality, we are constrained to using closed-source models (namely GPT and Claude) as they are the only ones currently able to reliably handle the Irish language.

3.1 Multi-Agent Architecture (Mul)

Our primary system instantiates a multi-agent framework in which all agents are implemented as calls to the GPT 4.1 model with role specific prompts. The workflow is defined as a directed graph with an orchestrator, one or more workers, a guardrail, and a finalizer that together control execution. To prevent unbounded recursive routing and reduce inference cost, the shared execution state maintains both a global iteration counter and per worker call limits, and the orchestrator halts or advances the pipeline once these bounds are reached; see overview in Figure 1.

Content Ordering Agent: The content ordering agent receives the full triple set X embedded in instructions from the Orchestrator and produces an ordered content plan (CO). The agent is instructed to group related facts, organise them into a coherent progression, and expose an ordering that later stages can follow. The resulting sequence serves as a high level content plan over the input triples.

Text Structuring Agent: The text structuring (TS) agent takes CO with sets of instructions from the orchestrator and converts it into a paragraph level outline which specifies paragraph boundaries and sentence level organisation for the document. This stage defines a document level skeleton that stabilises narrative flow before surface realisation.

Surface Realisation Agent: The surface realisation (SR) agent transforms the outline into fluent text in the target language ℓ. The agent is instructed to follow the outline closely, preserve the factual

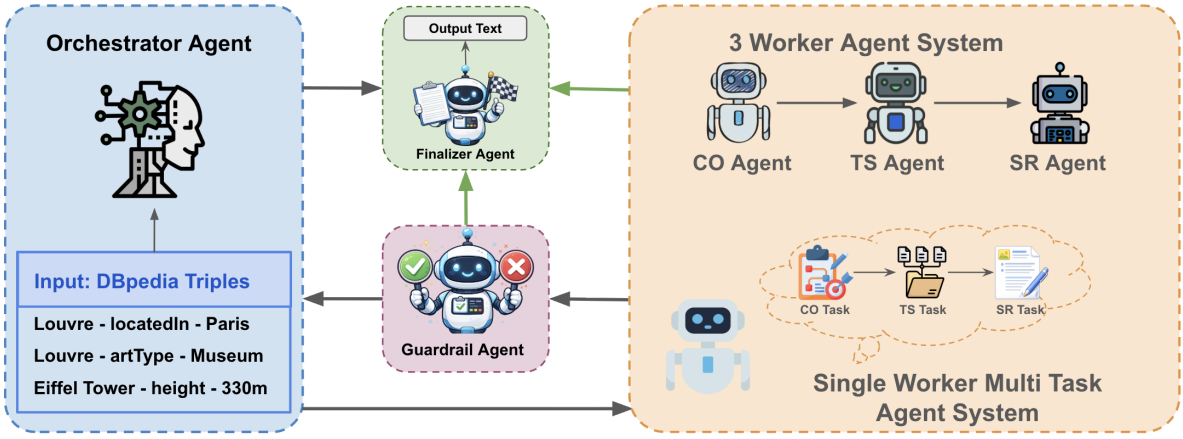

Figure 1: Multi-Agent System for Data-to-Text. CO = Content Ordering, TS = Text Structuring, SR = Surface Realization. Black arrows indicate the default flow. Green arrows indicate the near-limit shortcut: when execution reaches the penultimate global iteration $(n-1)$, the SR and Guardrail outputs are routed directly to the Finaliser, bypassing the normal path to preserve a usable output. The full algorithm is detailed in Algorithm 1.

content of each statement, and respect paragraph boundaries. We use language specific prompts for English and Irish that share the same structure but differ in examples and stylistic instructions. The agent does not have access to the raw triples, only to the structured outline and language, which encourages faithfulness to the planned content.

Orchestrator: The orchestrator coordinates the execution of agents. At each step it receives the current state S and the history of previous actions, then selects which agent to call next. When necessary, the orchestrator can request revisions of earlier stages, for example when the guardrail flags a violation or returns negative feedback.

Guardrail: The guardrail agent receives and evaluates intermediate outputs of the worker agents against task specific constraints and quality requirements. It checks that every triple in X is either mapped to a content plan CO or explicitly marked unused, that the TS outline obeys the required schema, that the SR draft text does not obviously contradict the input, and that no disallowed content is produced. When issues are detected, the guardrail generates structured feedback that the orchestrator uses to route the state back to the relevant worker with a targeted revision instruction.

Finalizer: The finalizer agent performs a last edit pass over the selected draft, standardising punctuation, resolving minor inconsistencies, and ensuring that the output meets quality guidelines. The finalizer cannot introduce new factual content.

3.2 Single Worker Multi-Agent Variant (Sng)

In addition to the three worker configuration (CO, TS, SR) described above, we also construct a single worker variant. This architecture uses a single worker agent that performs content ordering, text structuring, and surface realisation tasks within one prompt, still under the control of the same orchestrator, guardrail, and finalizer. This configuration allows us to examine whether explicit separation of roles is necessary, or whether a single multitask agent can approximate the same behaviour when embedded in the same control framework.

3.3 Baseline (e2e)

We define a simple end-to-end D2T system that uses a single GPT 4.1 LLM call per instance. Given an input triple set X and target language $\ell \in \{\text{en}, \text{ga}\}$, we pass the triples to GPT 4.1 together with a language-specific instruction prompt. This baseline does not expose intermediate representations and does not use an orchestrator, guardrail, or finalizer. Instead, it handles the entire task in a single forward pass, with temperature fixed to 0 for comparability with the multi-agent setting.

3.4 Dataset

In order to develop and test our models, we compiled a new dataset of DBpedia triples for a series of entities. We selected 900 random entities in total from three categories: *People*, *Geography* and *History*.[1] Following (Osuji et al., 2025a), for each

[1]The choice of categories is motivated by prior work (Belz et al., 2009) and by practical coverage considerations. These domains are the most densely populated and curated in DB-

entity, we retrieve all DBpedia triples in which it appears as Subject (e.g. **Einar_Leino** || **birthPlace** || Paltamo) or Object (e.g. Helkavirsiä || **author** || **Einar_Leino**), and curate the triple sets as follows: (a) filtering out properties identified as incorrect, including second to n^{th} occurrence of a Property that can only have one value (e.g. *birthDate*) and properties frequently misused on DBpedia, (b) allowing for a maximum of 6 instances of the same Property with the same Subject or Object,[2] and (c) selecting only the triple sets with at least 8 triples (i.e., triples sets of larger size than in WebNLG) and at most 199 triples (to keep the number of triples under control). Since all the collected triples have the queried entity as Subject or Object, the input graphs can be quite wide, but there is at most one level of chain relationships in this version of the dataset, i.e., the deepest subgraph would be: *Entity1* → *Property1* → *Queried Entity* → *Property2* → *Entity2*.

The resulting dataset comprises 669 input of size 8 to 199 triples each. We used 10 entities for developing the prompts, and randomly sampled 30 of the remaining entities for evaluation. We stratified the sampling by size so as to have 10 size bins (0-19, 20-39, etc.) and 3 data points per bin.

3.5 Evaluation

We ran the three systems (**Mul**, **Sng**, **e2e**) on all 669 datapoints and used the 30 sampled datapoints for the evaluation (98.37 triples/input on average). We use five criteria for the evaluation: `Grammaticality`, `Fluency`, `Coherence`, `No-Omissions` and `No-Additions`; see App. E for definitions as provided to human and LLM judges.

Human evaluation. Since it is extremely challenging to evaluate whether all and only input triples are being verbalised with large inputs,[3] we asked judges to rate intrinsic quality criteria only (`Grammaticality`, `Fluency`, `Coherence`). Outputs of different sizes and different systems were distributed as evenly as possible across six bilingual English/Irish speakers[4] via Latin Square assignment; see Appendix C for assignment details. Each output was evaluated on a 5-point scale by two annotators.

LLM-as-judge evaluation. For each input X and system output Y, we query an evaluator LLM (different from the LLM used for generating) with a structured rubric that asks for scores on the five criteria. The evaluator `claude-sonnet-4-5` receives X and Y, and returns numeric scores on the 1 to 5 scale in a json format. Since we have no human ratings for semantic accuracy, to test how sensitive the evaluator is to missing content in long texts, we create perturbed versions of the **e2e** output by deleting spans so that a proportion $\alpha \in \{0.05, 0.10, 0.15\}$ of the final sentences is removed. The three variants are identified as **e2e-0.95/0.90/0.85** respectively in the figures below. Similarly, to test how sensitive the evaluator is to added content, we create perturbed versions of the input (by removing a percentage of the final triples) paired with the full **e2e** texts. The three variants are identified as **e2e-input0.95/0.90/0.85** respectively in the figures.

4 Results and discussion

4.1 Human evaluation scores

Figure 2 shows the results of the human evaluation of the three main systems in English and Irish, with groupings based on Tukey's HSD post-hoc test (threshold 0.05).[5] The Multi-Agent system **mul** has consistently higher scores than the other two approaches across languages and criteria, particularly for `Coherence`, defined as the "degree to which an output's content/meaning hangs together better". However, almost none of the differences are statistically significant, possibly due to a small sample size. The results of the LLM-as-judge evaluation is more varied across criteria, but for `Coherence` the **mul** also scores higher in English.

4.2 LLM-as-judge scores

Table 4 in Appendix F summarises the LLM-as-judge scores, and Figures 5, 6 and 7 in Appendix H show visualisations of the LLM-as-judge evaluation. Figure 5 shows that for `No-Omissions`, `Fluency` and `Coherence`, all systems tend to get lower scores as the input size increases, especially the **e2e** system. This suggests that the **mul** and **sng** pipelines are somewhat more robust to

pedia, which makes it possible to extract large triple sets with sufficient volume and diversity for our analysis.

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

[3]In a pilot study, we carried out some tests with around 50 triples in the input and concluded that it was too challenging to manually evaluate semantic accuracy in a reliable way.

[4]Evaluators are all co-authors; one was involved in curating the dataset, the other five did only the evaluation; no evaluator knew anything about the evaluated systems.

[5]The scores of two systems who share a letter in the same table do not have statistically significant differences.

long inputs than a single pass prompt. At the same time, the ceiling scores of `No-Additions` and `Grammaticality` for both languages cast doubts on the ability of LLMs to evaluate these dimensions on long inputs, and point to an indifference to subtle hallucinations or grammatical differences in this scenario. However, Figures 6 and 7 show that LLMs seem to be able to detect missing and added content respectively, with the truncated outputs and inputs scores vertically ordered as expected for `No-Omissions` and `No-Additions`. Note that the difference between systems is clearer for `No-Additions` (Figure 7); for `No-Omissions` (Figure 6), the scores tend to converge for longer inputs (>100 triples), whereas we would still expect e.g. truncated outputs to score lower than full outputs. This could indicate that beyond a certain size, LLM-based assessment of missing contents may not be reliable. However, LLMs seem to be able to detect more accurately cases in which the generating systems produce additional contents not present in the input (usually referred to as *hallucinations* or *confabulations*).

(a) Grammaticality

	English		Irish	
	Mean	Group	Mean	Group
mul	**4.75**	A	**4.07**	A
e2e	4.70	A	3.97	A
sng	4.60	A	3.90	A

(b) Fluency

	English		Irish	
	Mean	Group	Mean	Group
mul	**4.32**	A	3.75	A
e2e	4.10	A	3.60	A
sng	4.07	A	**3.80**	A

(c) Coherence

	English		Irish	
	Mean	Group	Mean	Group
mul	**4.22**	A	**3.83**	A
e2e	4.02	AB	3.73	A
sng	3.85	B	3.60	A

Figure 2: System rankings human English and Irish

4.3 Correlation analysis

We first computed item-level Pearson and Spearman correlations between human scores, using all collected individual scores (we have two ratings per text by two different evaluators, and 90 texts/language). There is weak to moderate agreement on all criteria between human annotators on the English data (statistically significant Pearson correlations in the 0.3–0.5 range for all three criteria) but negative or near-zero Spearman's and Pearson's correlations on the Irish data, which are very likely due to major rater bias differences, with three of the six annotators having very different patterns (lenient, harsh or lenient on some criteria and harsh on some others). Since the same annotators evaluated English and Irish texts, we believe that this highlights classic low-resource language challenges: raters may disagree on what "good" grammar/fluency looks like due to dialect variation, normalization, or lack of standards.

We also computed item-level Pearson and Spearman correlations between averaged human scores and LLM-as-judge scores (90 texts/language). There is a weak, non-significant LLM-human alignment in both English and Irish (non statistically significant 0.1 Pearson's correlation for two criteria). For `Grammaticality`, it was not possible to compute correlations because there is no variance for the LLM scores. Given the lack of correlation between human ratings in Irish, it is difficult to interpret these human-LLM correlation numbers. But the case of English hints that LLMs may not be reliable for the evaluation of large input/output pairs beyond a certain size (>100 triples).[6]

5 Conclusion

We have explored the utility of a multi-agent approach for the generation of long texts in English and Irish. Our human evaluation results show a tendency of multi-agents to score higher than a single end-to-end system, but the differences were not statistically significant. This is in line with our experimental design, where all systems share the same strong backbone model and differ mainly in control flow rather than in underlying language ability. Although multi-agent generation requires multiple LLM calls and therefore increases computational cost compared with a single end-to-end pass, our findings still indicate that the approach is promising. The results show slightly superior coherence for the multi-agent approach in both languages, yet weak/moderate (English) or absent (Irish) human inter-rater agreement for all criteria, and very weak human-LLM alignment overall, exposing key limits in scalable NLG evaluation. Finally, we observe a systematic gap between English and Irish in human ratings but not in the LLM-as-judge setting. This suggests that multilingual long-context generation and evaluation remain challenging, particularly for lower-resource settings.

[6] See Appendix G for system-level correlations.

Limitations

Use of closed-source models. We are not using open-weight models in our experiments because none is currently able to produce reliable texts in Irish language; we are restricted to using either GPT or Claude. We will explore open-weight models as they become reliable in an under-resourced language such as Irish.

Computational cost. The proposed multi-agent approaches have a high computational cost due to the multiple calls to large language models. However, the objective of our experiments is to assess whether or not it is worth pursuing along the lines of multi-agent D2T generation, and we leave the reduction of the resource use as future work.

Data quality. Despite our filters to ensure a high quality in the input triple sets, a small percentage of triples encode "wrong" information that impact the text quality, in particular in terms of semantic content: e.g. *Ibn al-Tilmidh – occupation – Baghdad* is a triple in which there should be a job title instead of "Baghdad". The resulting text generated by the systems can be confusing to evaluators, and impact both *Fluency* an *Coherence* scores.

Results impact. The human evaluation is limited to 30 texts per system and per language, and we did not carry out ablation studies, because of the high cost of evaluating long text quality, even focusing on intrinsic quality criteria such as `Grammaticality`, `Fluency` and `Coherence`. Evaluators spend between 5 and 10 hours each for evaluating about 60 texts (equally distributed between English and Irish). With such a limited sample size, it is difficult to detect statistically reliable differences between similar systems (according to Tukey HSD with a threshold of 0.05), even though we consistently observe directional trends towards improved coherence with the modular approach. We expect that a larger annotated sample could confirm these trends, and future human evaluations would need to be carried out on a larger scale, if it is indeed possible (evaluator recruitment for under-resourced languages is very challenging).

Ethics Statement

We use LLM-based methods in our experiments, and at present, it is uncertain what data has been used to train them, especially proprietary models such as GPT and Claude. The texts they produced and the assessments they provided may reflect biases, potentially posing a risk of harm to users. All human evaluators are co-authors of the paper and have not been involved in the development of either the generating systems or evaluation methods. We used LLM assistance for rewriting occasional sentences and checking some of the code used to analyse the results.

Acknowledgments

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

The ADAPT members' contribution was funded by the Irish Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media via the eSTÓR project. Our work has also benefited more generally from being carried out within the research environment of the ADAPT SFI Centre, funded by Science Foundation Ireland through the SFI Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106_P2.

References

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

Anya Belz, Simon Mille, and Craig Thomson. 2025. The qcet taxonomy of standard quality criterion names and definitions for the evaluation of nlp systems.

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.

Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. 2025. A survey on the optimization of large language model-based agents. *arXiv preprint arXiv:2503.12434*.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2025. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7376–7399.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.

Mateusz Lango and Ondřej Dušek. 2025. Llm agents implement an nlg system from scratch: Building interpretable rule-based rdf-to-text generators. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*.

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

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

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024b. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

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

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

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024b. Longgenbench: Long-context generation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 865–883.

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

Chinonso Cynthia Osuji, Simon Mille, Ornait O'Connell, Thiago Castro Ferreira, Anja Belz, and Brian Davis. 2025a. Scaling up data-to-text generation to longer sequences: A new dataset and benchmark results for generation from large triple sets. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 810–822.

Chinonso Cynthia Osuji, Brian Timoney, Mark Andrade, Thiago Castro Ferreira, and Brian Davis. 2025b. Are multi-agents the new pipeline architecture for data-to-text systems? In *Proceedings of the 18th International Natural Language Generation Conference*, pages 542–553.

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

Kaiyang Wan, Honglin Mu, Rui Hao, Haoran Luo, Tianle Gu, and Xiuying Chen. 2025. A cognitive writing perspective for constrained long-form text generation. *arXiv preprint arXiv:2502.12568*.

Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. 2025. Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems. *arXiv preprint arXiv:2502.11098*.

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

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Juanzi Li, and Roy Ka-Wei Lee. 2025a. Superwriter: Reflection-driven long-form generation with large language models. *arXiv preprint arXiv:2506.04180*.

```
<entry category="people" eid="360" shape="(X (X) (X) (X) (X))" shape-type="sibling" size="9">
  <originaltripleset>
    <otriple>Eino_Leino | birthPlace | Paltamo</otriple>
    <otriple>Eino_Leino | deathPlace | Tuusula</otriple>
    <otriple>Eino_Leino | birthDate | 1878-07-06</otriple>
    <otriple>Eino_Leino | birthName | Armas Einar Leopold Lönnbohm</otriple>
    <otriple>Eino_Leino | birthYear | 1878</otriple>
    <otriple>Eino_Leino | deathDate | 1926-01-10</otriple>
    <otriple>Eino_Leino | deathYear | 1926</otriple>
    <otriple>Helkavirsiä | author | Eino_Leino</otriple>
    <otriple>L._Onerva | partner | Eino_Leino</otriple>
  </originaltripleset>
</entry>
```

Figure 3: Sample input triple set of size 9 (small triple set for better readability)

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

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Ruibin Xiong, Yimeng Chen, Dmitrii Khizbullin, Mingchen Zhuge, and Jürgen Schmidhuber. 2025. Beyond outlining: Heterogeneous recursive planning for adaptive long-form writing with language models. *arXiv preprint arXiv:2503.08275*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

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

Algorithm 1 Multi-agent LLM framework for data-to-text generation

Require: Structured input X (triples), target language ℓ, configuration θ
Ensure: Document level text Y

$S \leftarrow \textsc{InitState}(X, \ell, \theta)$
$H \leftarrow \varnothing$
$k \leftarrow 0$
while not $\textsc{Converged}(S)$ **and** $k < \theta$.max_iterations **do**
  $k \leftarrow k + 1$
  if $k = \theta$.max_iterations **and** S.output **is not** set **then**
    $r \leftarrow \textsc{Finalization}$
  else
    $r \leftarrow \textsc{Orchestrator}(S, H)$
  if $r = \textsc{ContentOrdering}$ **then**
    $C \leftarrow \textsc{ContentOrderingAgent}(X, S)$
    S.content_plan $\leftarrow C$
  else if $r = \textsc{TextStructuring}$ **then**
    $O \leftarrow \textsc{TextStructuringAgent}(S$.content_plan$, S)$
    S.outline $\leftarrow O$
  else if $r = \textsc{SurfaceRealization}$ **then**
    $D \leftarrow \textsc{SurfaceRealizationAgent}(S$.outline$, \ell, S)$
    S.draft $\leftarrow D$
  else if $r = \textsc{Guardrail}$ **then**
    $(F, G) \leftarrow \textsc{GuardrailAgent}(X, S)$
    S.feedback $\leftarrow F$
    S.guardrail_status $\leftarrow G$
  else if $r = \textsc{Finalization}$ **then**
    $Y \leftarrow \textsc{FinalizerAgent}(S$.draft$, S)$
    S.output $\leftarrow Y$
  $H \leftarrow \textsc{UpdateHistory}(H, r, S)$
  $S \leftarrow \textsc{UpdateState}(S, r)$
if S.output **is not** set **then**
  $Y \leftarrow \textsc{FallbackGenerate}(X, \ell, \theta)$
else
  $Y \leftarrow S$.output
return Y

Appendix

A Sample input

Figure 3 shows a sample input used to generate the texts in both English and Irish. The input is of size 9 (9 triples); we have inputs of up to size 199.

B Algorithms

Algorithm 1 details the steps of the multi-agent framework.

C Details on Latin square assignments

In order to maximise the distribution of annotators across systems and data points, since we have 3 different systems and 2 annotators per data point, pairs of annotators [Annotator N, Annotator N+2] were assigned the same evaluation item when possible. So for instance, if we consider the first 3 (out of 30) data points, System 1 has the following respective pairs of annotators: [[A1, A3], [A2, A4], [A3, A5]]; System 2 has [[A2, A4], [A3, A5], [A4, A6]] respectively, and System 3 has [[A3, A5], [A4, A6], [A1, A5]] respectively. This ensures that (i) each system is seen by 5 different annotators, and (ii) each datapoint is seen by 5 different annotators. The code for Latin square design assignment and automatic compilation of evaluation spreadsheets for individual evaluators is available publicly: https://github.com/mille-s/Build_KGs_entities.

D LLM-as-judge evaluation

We model the evaluator as a scoring function $J_d(X, Y) \in \{1, 2, 3, 4, 5\}$, where d indexes the evaluation criterion, X is the triple set, and Y is the system output. Let there be N evaluator LLMs, in our case $N = 1$. For a given system θ and an evaluation set $\mathcal{D}_{\text{eval}}$, the average score for criterion d is

$$\bar{J}_d(\theta) = \frac{1}{N\,|\mathcal{D}_{\text{eval}}|} \sum_{m=1}^{N} \sum_{(X,Y_\ell)\in\mathcal{D}_{\text{eval}}} J_d^{(m)}\big(X,\, \hat{Y}_\theta(X,\ell),\, Y_\ell\big), \quad (1)$$

where $\hat{Y}_\theta(X, \ell)$ denotes the output generated by system variant θ in language ℓ, and $J_d^{(m)}$ is the score assigned by evaluator m on criterion d. These per criterion averages $\bar{J}_d(\theta)$ are the scores we report.

Here θ indexes a particular system variant, for example the three worker multi-agent model, the unified worker model, or the end-to-end baseline.

E Evaluation dimensions

The criteria below were named and defined following the quality criteria names and definitions in the QCET Taxonomy (Belz et al., 2025), for comparability and reproducibility.

E.1 Instructions provided to human evaluators

Human evaluators carried out the assessment in Google Spreadsheets which contained the instructions as described in this section. Below the instructions, each row contained a text and three cells with drop-down menus that allow for selecting one of the five ratings. Each evaluator was provided with on sheet for English texts and one sheet with Irish texts, with the texts on the same row corresponding to the same input. All released evaluations will be fully anonymised via random ID assignment to each evaluator.

`Grammaticality`: To what degree is this text free of grammatical errors, looking at its form only? Grammaticality exclusively captures syntactic correctness, ignoring its content/meaning.

- Very good: There are no grammatical errors in the text.
- Good: Somewhere between Very good and Fair.
- Fair: There are a few grammatical errors in the text.
- Poor: Somewhere between Very poor and Fair.
- Very poor: There are numerous grammatical errors in the text.

`Fluency`: To what degree is this output fluent? Fluency captures how well the text flows, and can be absorbed readily without bringing the reader up short.

- Very good: The text flows well and can be read easily.
- Good: Somewhere between Very good and Fair.
- Fair: The text has occasional disfluencies.
- Poor: Somewhere between Very poor and Fair.
- Very poor: The text does not flow well and I had to start over to understand some parts.

`Coherence`: To what degree does this output's content/meaning hang(s) together better? The text should be well-structured and well-organized from the perspective of meaning only, i.e. without taking into account the quality of the form. The text contents should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

- Very good: The text is well organised and coherent.
- Good: Somewhere between Very good and Fair.
- Fair: The text could be better organised for a better coherence.
- Poor: Somewhere between Very poor and Fair.
- Very poor: The text is poorly organised and lacks coherence.

E.2 Prompt for LLM-as-judge evaluation

In the prompt, the triples are formatted as simple sequences of "Subject Property Object;", with no separators between the components of each triple.

prompt = "' 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 five specific Dimensions to evaluate against:

Dimensions:""" `No-Omissions`: To what degree is ALL the information in the Triple Set present in the Text? `No-Additions`: To what degree is ONLY information from the Triple Set present in the Text? `Grammaticality`: To what degree is the Text free of grammatical errors, looking at its form only? `Coherence`: To what degree is the Text well-structured and well-organized into coherent body of information about a topic, from the perspective of meaning only? `Fluency`: To what degree does

the Text flow, and can be absorbed readily without bringing the reader up short?"""

Important note on `No-Omissions` and `No-Additions`: 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`, `Coherence` and `Fluency`: for `Grammaticality`, `Coherence` 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 5 (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": "", "Coherence": "Justification": "", "Score": "", "Fluency": "Justification": "", "Score": "" .

Make sure to read thoroughly the Triple Set and the "'+str(VARIABLE-Language)+"' Text below, and assess the five Dimensions using the instructions and template above.

Triple Set: "' + str(VARIABLE-Triples) + "<linebreak>" + "'Text: "'+ str(VARIABLE-Text) + "<linebreak><linebreak>" + "' "'

F LLM-as-judge scores

Figure 4 shows the LLM-as-judge evaluation results.

G System-level correlations between Human and LLM-as-judge scores

We computed system-level Pearson and Spearman correlations between human and LLM ratings for each criterion (average of both human raters for each system VS average LLM scores for each system). In English, the systems are ranked the same for `Coherence` (Spearman r = 1), and the score difference between the systems is also similar (Pearson r = 0.98). For `Fluency`, the Spearman rank correlation is 0.5, and Pearson is 0.95. For Irish, both Pearson and Spearman correlations are largely negative for both criteria. These numbers are based on a very small sample, so the p-values are not very informative. No correlations could be computed for `Grammaticality` because of the lack of variance of the LLM scores.

(a) No-omissions

	English		Irish	
	Mean	Group	Mean	Group
sng	**4.80**	A	**4.67**	A
mul	4.60	A	4.43	AB
e2e	4.57	A	4.40	AB

(b) No-additions

	English		Irish	
	Mean	Group	Mean	Group
sng	**4.87**	A	**5.00**	A
mul	4.77	A	**5.00**	A
e2e	4.53	A	4.87	A

(c) Grammaticality

	English		Irish	
	Mean	Group	Mean	Group
sng	5.00	A	5.00	A
mul	5.00	A	5.00	A
e2e	5.00	A	5.00	A

(d) Fluency

	English		Irish	
	Mean	Group	Mean	Group
mul	**4.30**	A	4.30	A
sng	4.23	A	4.50	A
e2e	4.07	A	**4.73**	A

(e) Coherence

	English		Irish	
	Mean	Group	Mean	Group
mul	**4.80**	A	4.77	A
e2e	4.60	A	**4.90**	A
sng	4.50	A	**4.90**	A

Figure 4: System rankings LLM

H Plots

Figures 5, 6 and 7 show the plots for LLM-as-judge evaluation across the different input sizes. On Figure 6, as expected, for `No-Omissions` the **e2e** curve is higher than the **e2e-0.95** curve, which is higher than the **e2e-0.90**, which is higher than the **e2e-0.85**. Note that with longer inputs in Irish the difference is not so clear. Similarly, in Figure 7, for `No-Additions` the curves for the texts with the most truncated inputs are lower on the plot in the expected order, but in a clearer way than it was the case for `No-Omissions`. LLMs seem better at assessing `No-Additions` (i.e. hallucinations) than `No-Omissions` (i.e. missing content).

I Qualitative Comparison of End-to-End and Pipeline Outputs

To complement the quantitative evaluation, we manually inspected paired outputs from the end-to-end baseline and the default pipeline system. The clearest differences arise when the input combines heterogeneous fact types, such as locations, his-

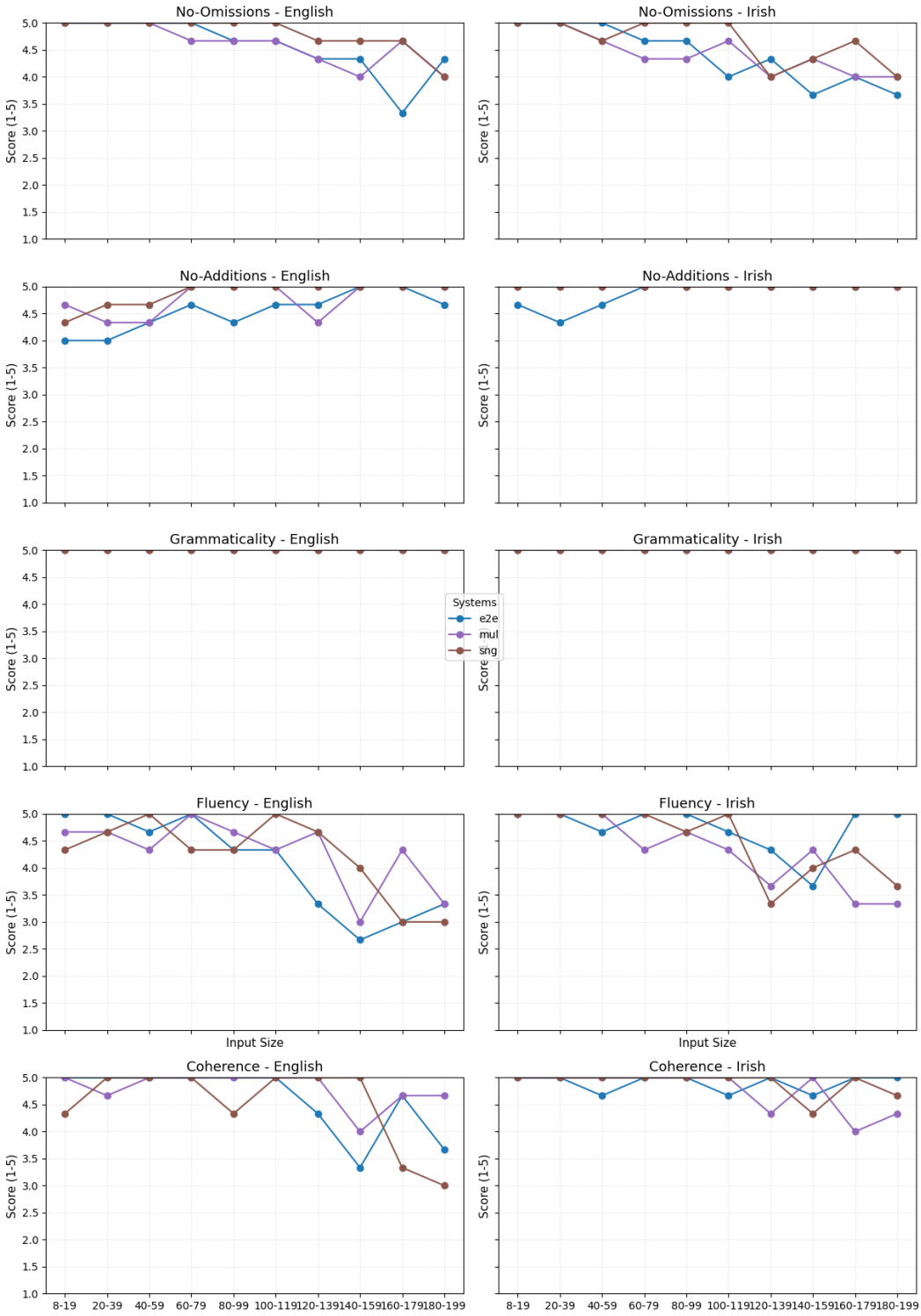

Figure 5: System performance across criteria, input sizes and languages

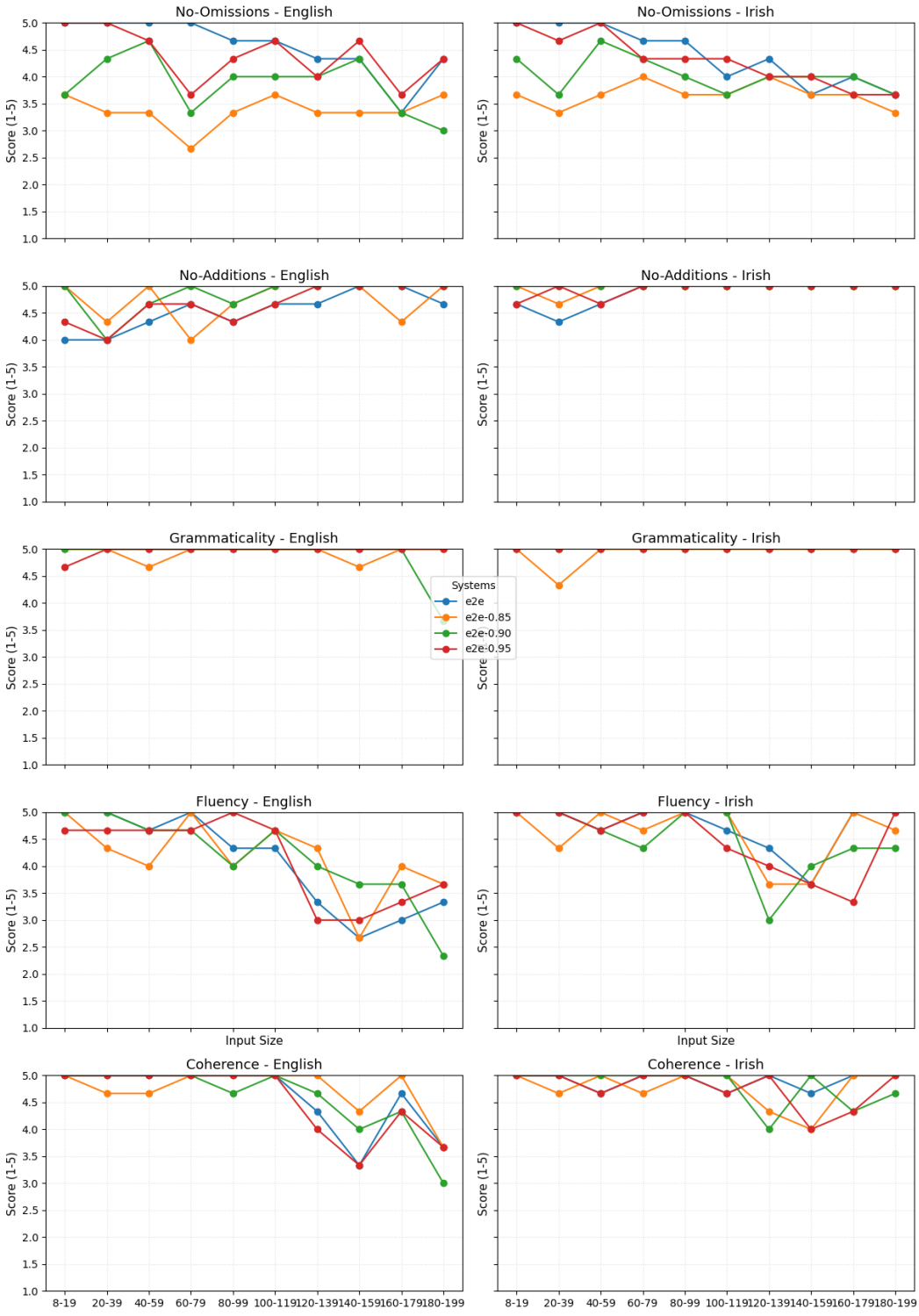

Figure 6: E2E System performance across criteria, input sizes and languages with truncated outputs (shows LLM sensitivity to missing contents).

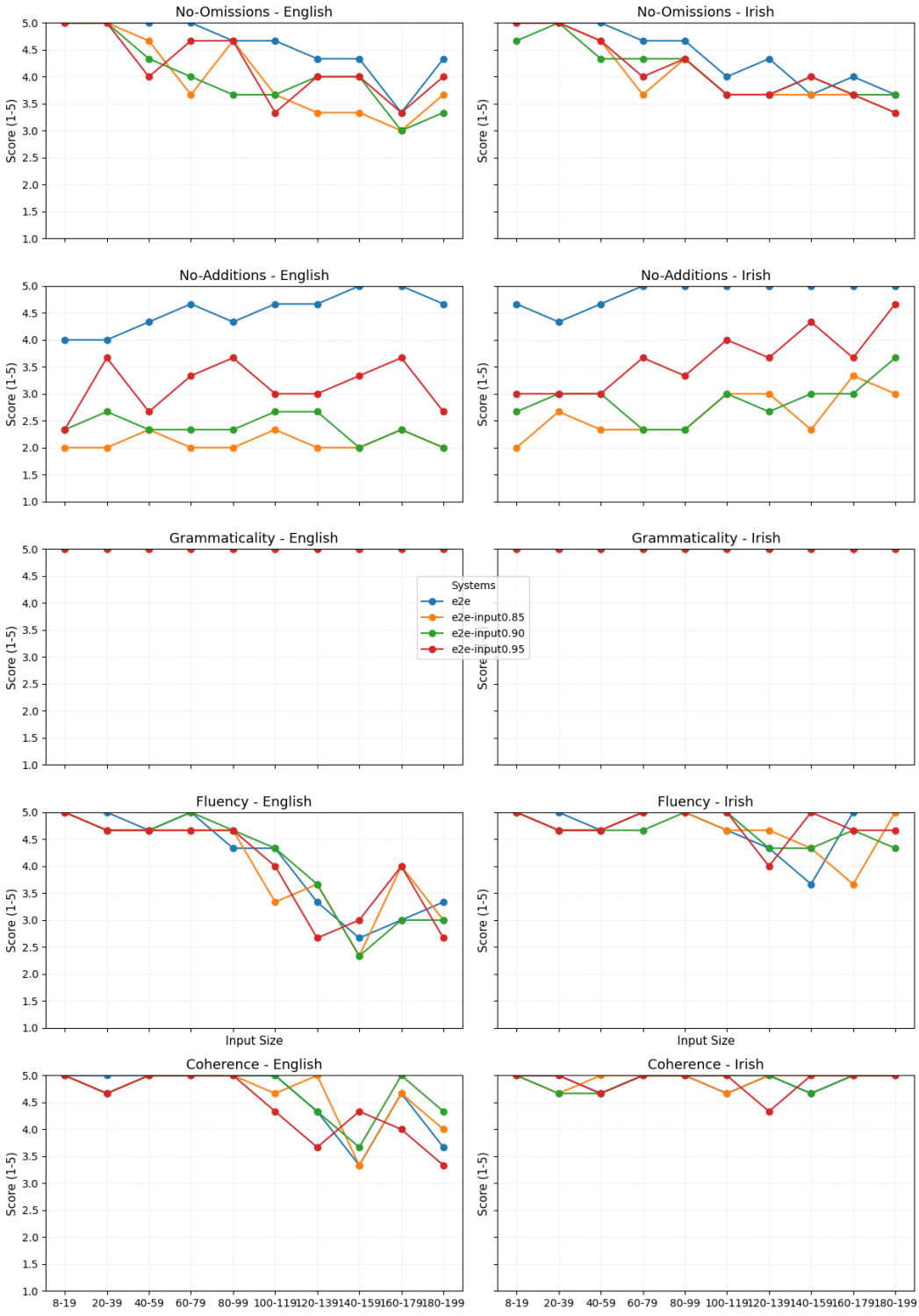

Figure 7: System performance across criteria, input sizes and languages with truncated inputs (shows LLM sensitivity to added contents).

torical events, population facts, and person-related information. In such cases, the pipeline more often groups related facts into topical clusters before realisation, whereas the end-to-end system more often opens with a broad introductory sentence that is not directly grounded in any single input triple, though it may loosely characterise the set as a whole.

The strongest example is *Atacama Desert*. The pipeline organises the content into a clearer progression: it begins with the subdivision facts, then introduces the geographic landmark, followed by the population and language facts, and ends with the person-related facts. The end-to-end output, by contrast, opens with a general framing clause. The opening sentence, "*The Atacama Desert is associated with a variety of notable people, places, and cultural elements*", appears before presenting person, place, and cultural information ahead of the subdivisions. While this opening is loosely inferable from the triple set as a whole, it is not directly grounded in any single triple, and the subsequent ordering is less cohesive than the pipeline's.

A similar pattern appears in *Altiplano*. The end-to-end output opens with "*The Altiplano is a region that serves as the location for several notable geographical and historical features*", a framing clause that compresses the remaining information into a single descriptive unit rather than presenting the facts in a structured progression. The pipeline output is more direct and content-led: it groups the lakes first, then the mountain and regional facts, and closes with the historical and population-related information, making the ordering of facts more transparent.

The *Balkan Mountains* example shows the same tendency. The end-to-end output opens with "*The Balkan Mountains serve as the location for several notable sites and events*", then mixes location and battle-related facts without clear topical separation. The pipeline instead presents the mountain peaks first, then the other located entities, and finally the historical facts, yielding a clearer category-based progression.

Across these three examples, the introductory framing sentences produced by the end-to-end system share a common structure: they are broad summaries that are weakly derivable from the triple set as a whole but not explicitly supported by any individual triple. Whether these constitute hallucinations in the strict sense is debatable, since they do not introduce factually incorrect information. They are better characterised as over-generalisations that substitute a vague descriptor for the specific structured content the input encodes. The pipeline avoids this pattern entirely by beginning directly with the most specific facts.

These examples highlight the instance-level benefit that motivates our hypothesis: modular planning can improve discourse organisation by grouping related facts before final realisation. The pipeline consistently shows clearer topical grouping, more direct content-led ordering, and fewer unsupported framing statements. This interpretation is consistent with the broader evaluation in the main paper, where modular systems show their clearest advantage on coherence, even though corpus-level differences remain modest. Representative paired outputs are presented in Table 1.

Input Triples (condensed)	End-to-End Output	Pipeline Output
`Tom_Marsh | deathPlace` `Guillermo_Chong | knownFor` `Cerro_Paranal | locatedInArea` `Chango_people | populationPlace` `Kunza_language | spokenIn` *Subject: Atacama_Desert (+ 5 subdivision triples)*	*The Atacama Desert is associated with a variety of notable people, places, and cultural elements.*† Tom Marsh, an astronomer, passed away in the Atacama Desert, while GuillermoChong is known for his work related to this region. Cerro Paranal, a significant location, is situated within the Atacama Desert. The Chango people have their population in the Atacama Desert, and the Kunza language is spoken in this area. Several subdivisions are part of the Atacama Desert, including Alto del Carmen, Caldera, Chañaral, Copiapó, and Diego de Almagro in Chile.	The Atacama Desert includes several subdivisions, such as Alto del Carmen, Caldera in Chile, Chañaral, Copiapó, and Diego de Almagro in Chile. Cerro Paranal is located within the Atacama Desert. The Chango people have lived in this region, and the Kunza language has been spoken there. Guillermo Chong is known for his association with the Atacama Desert, while Tom Marsh, the astronomer, died in this area.
`Laguna_Verde | location` `Lake_Poopó | location` `Colla-Inca_War | place` `Lake_Tauca | isPartOf` `Cerro_Tujle | mountainRange` *Subject: Altiplano (+ 3 more triples)*	*The Altiplano is a region that serves as the location for several notable geographical and historical features.*† Laguna Verde and Lake Poopó are both situated within the Altiplano, and Lake Tauca is also part of this region. The Altiplano is home to mountain ranges such as Cerro Tujle and the Lauca volcano, as well as the Cordillera Real in Bolivia. Historically, the Altiplano was the site of the Colla–Inca War. The region is also significant for its population of indigenous peoples in Bolivia.	Laguna Verde in Bolivia and Lake Poopó are both located in the Altiplano region, and Lake Tauca is also part of the Altiplano. Cerro Tujle and the Lauca volcano are mountains within the Altiplano mountain range, while the Cordillera Real in Bolivia is a region associated with the Altiplano. The Altiplano is home to indigenous peoples in Bolivia and was also the site of the Colla–Inca War.
`Stara_Planina_WF | location` `Sinite_Kamani_NP | location` `Tsonevo_Reservoir | location` `Battle_Rishki_Pass | place` `Mount_Chepan | mountainRange` *Subject: Balkan_Mountains (+ 9 more triples)*	*The Balkan Mountains serve as the location for several notable sites and events.*† The Stara Planina Wind Farm, Sinite Kamani Nature Park, Tsonevo Reservoir, Central Balkan National Park, and Uzana in Bulgaria are all situated within this mountain range. Additionally, the Bebresh Viaduct is located in the area of the Balkan Mountains. Several significant battles took place here, including the Battle of the Rishki Pass, the Battle of Zlatitsa, and the Battle of Eski Zagra. The mountain range also encompasses prominent peaks such as Mount Chepan, Botev Peak, Buzludzha, Midžor, and Levski Peak in Bulgaria.	Mount Chepan, Botev Peak, Buzludzha, Midžor, and Levski Peak in Bulgaria are all part of the Balkan Mountains. The Stara Planina Wind Farm, Sinite Kamani Nature Park, Tsonevo Reservoir, Central Balkan National Park, and Uzana in Bulgaria are located in the Balkan Mountains, as is the Bebresh Viaduct. The Balkan Mountains were also the site of the Battle of the Rishki Pass and the Battle of Zlatitsa, and served as the territory for the Battle of Eski Zagra.

Table 1: Paired end-to-end and pipeline outputs for three representative examples. Input triples are condensed for space; full triple sets and complete outputs for all evaluated instances are available in the project repository. Sentences marked with † are broad introductory generalisations that are not directly grounded in any single input triple, though they are loosely inferable from the triple set as a whole. They are better characterised as over-generalisations than strict hallucinations, since they do not introduce factually incorrect information.