Exploring the impact of data representation on neural data-to-text generation

David M. Howcroft and **Lewis Watson** and **Olesia Nedopas** and **Dimitra Gkatzia**
School of Computing, Engineering, and the Built Environment
Edinburgh Napier University
Edinburgh, Scotland, United Kingdom
{d.howcroft,l.watson,o.nedopas,d.gkatzia}@napier.ac.uk

Abstract

A relatively under-explored area in research on neural natural language generation is the impact of the data representation on text quality. Here we report experiments on two leading input representations for data-to-text generation: attribute-value pairs and Resource Description Framework (RDF) triples. Evaluating the performance of encoder-decoder seq2seq models as well as recent large language models (LLMs) with both automated metrics and human evaluation, we find that the input representation does not seem to have a large impact on the performance of either purpose-built seq2seq models or LLMs. Finally, we present an error analysis of the texts generated by the LLMs and provide some insights into where these models fail.

1 Introduction

In the field of Natural Language Generation (NLG), the quality of generated text is crucial, influencing the usability and effectiveness of applications ranging from automated reporting to conversational agents. The focus of the field has predominantly been on developing more sophisticated models and algorithms creating a gap in understanding the impact of input data representations. Over the years, various input representations for end-to-end NLG have been utilised. These representations have often been chosen based on convenience, such as pre-existing formats of input data or prevailing trends. However, to our knowledge, no previous research has systematically investigated whether the choice of input representation affects the overall quality of the generated text. By addressing this gap, our study aims to evaluate how different input representations impact the fluency and semantic fidelity of generated texts. This investigation not only contributes to theoretical advancements but also offers practical insights into improving NLG systems.

NLG systems utilise various input representations to convert structured data into text. These representations include attribute-value pairs, as in the End-to-End Generation Challenge (Dušek et al., 2020, E2E), where each pair provides specific details about an entity, such as a restaurant's name, type, cuisine, price range, customer rating, and location. Another popular format is Resource Description Framework (RDF) triples, exemplified by the WebNLG dataset (Gardent et al., 2017), where each input consists of a subject-predicate-object structure, enabling the system to generate text based on relationships between entities, such as 'Edinburgh is the capital of Scotland'.

In this paper, we explore the impact of input representations in data-to-text generation, i.e. in tasks where the input of an NLG system is structured data and the output is coherent and contextually relevant natural language texts. We explore the classic

E2E

```
name == Blue Spice <PAIR_SEP> eat type == coffee shop <PAIR_SEP> area == city centre
```

Blue Spice is a coffee shop located in the city centre.

WebNLG

```
<SUBJECT> Above the Veil <PREDICATE> number of pages <OBJECT> 248 <TRIPLE_SEP> <SUBJECT> Above the Veil <PREDICATE> author <OBJECT> Garth Nix <TRIPLE_SEP> <SUBJECT> Above the Veil <PREDICATE> media type <OBJECT> Hardcover
```

"Above the Veil" by Garth Nix is a 248-page hardcover book.

Figure 1: Two linearistations of E2E and WebNLG inputs. E2E's input format consists of attribute-value pairs. WebNLG's inputs are semantic triples, composed of subject, predicate and object.

Proceedings of the 17th International Natural Language Generation Conference, pages 243–253
September 23–27, 2024.

seq2seq NLG architecture (exemplified by (Dušek and Jurčíček, 2016)) and Large Language Models (LLMs; in particular, GPT (OpenAI et al., 2024) and Llama (Touvron et al., 2023)) with two popular tasks and their corresponding input formats, namely E2E and WebNLG. In order to represent these input formats as sequences for neural network models, we linearise them as shown in Figure 1.

This paper examines the following research question: 'Do input representations matter in data-to-text systems?'. Our contributions are: (1) we present a comparison of two leading representations for data-to-text research for neural seq2seq models and LLMs; and (2) we provide the code for reproducing these experiments with other linearisations of comparable meaning representations at `https://github.com/NapierNLP/inlg2024`.

Our careful human evaluations across two datasets find no statistically significant evidence that attribute-value representations or RDF representations are superior across the board. Comparing trends within a single system, our results suggest that there may be a slight benefit of using RDFs for accuracy for Llama 3 or for seq2seq models, with a slight penalty to fluency, though further research is necessary given the small differences in performance on these datasets. A qualitative error analysis confirms that GPT-4o and Llama 3 produce very few semantic errors in these domains, though Llama 3 does sometimes omit content from more complicated RDF inputs and both can produce occasionally stilted language.

2 Datasets

We adopt the enriched versions of the WebNLG and E2E datasets, since they have both been prepared similarly from existing datasets. For our work, we limit ourselves to using the raw inputs and outputs and corresponding delexicalisations.

For the Enriched WebNLG dataset, Castro Ferreira et al. (2018) adapt the WebNLG corpus to include annotations for content ordering, sentence segmentation, surface realisation, and referring expression generation (REG). Delexicalisation was performed manually, labelling the subjects for RDF predicates as AGENTs and the objects as PATIENTs, with numeral suffixes to indicate which predicate the entities are associated with. Entities which appear in both subject and object roles for different predicates in the same input are delexicalised with the label BRIDGE.

For the Enriched E2E dataset, Castro Ferreira et al. (2021) adapt the E2E Challenge corpus (Novikova et al., 2017) to include annotations for content ordering, sentence segmentation, lexicalisation, REG, and surface realisation. Where the Enriched WebNLG dataset treated lexicalisation and surface realisation in a single step, with REG as a post-process, the Enriched E2E dataset handles lexicalisation and surface realisation separately.

Linearisation We process the raw XML files provided for the two datasets to create the linearisation for each input. For WebNLG, we extract each RDF triple and render its component subject, predicate, and object in sequence, preceded by a label in angled brackets. Between each triple, we insert a `<TRIPLE_SEP>` label as a separator. For E2E, each attribute-value pair is linearised as `attribute == value`, with the label `<PAIR_SEP>` separating each pair from the next. All underscores were replaced by space characters and any `camelCase` text was rendered instead as sequences of space-separated words (i.e. `camel case`). For example, the original XML representations for the inputs shown in Figure 1 are shown in Figure 2.

3 Models

We explore a classic approach to neural data-to-text generation as well as zero-shot LLM prompting for this work.

Seq2Seq+Attn TGen (Dušek and Jurčíček, 2016) is the seq2seq model with attention which was used as a baseline for the End-to-End Challenge (Dušek et al., 2020) and remains a competitive baseline for data-to-text tasks. We adapt the reimplementation from Howcroft and Gkatzia (2023), which uses PyTorch instead of Tensorflow and uses more up-to-date dependencies, to work with our task where the inputs do not have to be in the exact format expected by TGen. This model omits the semantic error reranker from TGen.

Open and Closed LLMs For LLMs we explored two recently released models, one open (Llama 3) and one proprietary (GPT-4o).[1] The open model is our priority, as model availability is essential to reproducibility and inspectability, but GPT-4o is included as it represents the latest advancements in proprietary language models. The

[1]There are no technical reports for either model yet; however, the Model Card for Llama 3 is available: AI@Meta (2024).

E2E

```
<input attribute="name" tag="__NAME__" value="Blue Spice"/>
<input attribute="eatType" tag="__EATTYPE__" value="coffee shop"/>
<input attribute="priceRange" tag="__PRICERANGE__" value="£20-25"/>
<input attribute="customer rating" tag="__CUSTOMER_RATING__" value="3 out of 5"/>
<input attribute="area" tag="__AREA__" value="city centre"/>
<input attribute="familyFriendly" tag="__FAMILYFRIENDLY__" value="no"/>
<input attribute="near" tag="__NEAR__" value="Avalon"/>
```

```
name == Blue Spice <PAIR_SEP> eat type == coffee shop <PAIR_SEP> price range ==
£20-25 <PAIR_SEP> customer rating == 3 out of 5 <PAIR_SEP> area == city centre
<PAIR_SEP> family friendly == no <PAIR_SEP> near == Avalon
```

WebNLG

```
<mtriple>Above_the_Veil | numberOfPages | “248”</mtriple>
<mtriple>Above_the_Veil | author | Garth_Nix</mtriple>
<mtriple>Above_the_Veil | mediaType | Hardcover</mtriple>
```

```
<SUBJECT> Above the Veil <PREDICATE> number of pages <OBJECT> 248 <TRIPLE_SEP>
<SUBJECT> Above the Veil <PREDICATE> author <OBJECT> Garth Nix <TRIPLE_SEP>
<SUBJECT> Above the Veil <PREDICATE> media type <OBJECT> Hardcover
```

Figure 2: Enriched E2E and WebNLG corpora inputs corresponding to the examples shown in Figure 1, with our linearisations repeated here for convenience.

System prompt *You are a linguistic robot that translates messages from an input data format into text.*
User prompt *Perform data-to-text generation using the following data. Be concise. Do not include any other information.*

Table 1: Prompts used for GPT-4o and Llama 3

prompting was done through Unify[2], a service providing access to a variety of LLMs. For this research, we used Llama 3 with 70B parameters. The total cost of running these experiments was 12.50 USD through Unify.

Each entry from the datasets was sent to both models along with a system and user prompts, which are shown in Table 1. This prompt was chosen after testing 10 different prompts across both datasets with GPT-4o.

4 Automatic Evaluations

We use reference-based automated metrics primarily to assess the degree to which our seq2seq model learns to match the kinds of texts present in the corpora, though we also report the LLMs' performance for reference. We report BLEU (Papineni et al., 2002) as implemented in SacreBLEU[3] (Post, 2018) for a discrete word-overlap metric and rescaled BERTScore[4] F1 (Zhang et al., 2020) for a slightly more flexible quality metric.

Table 2 shows the results for the E2E Challenge dataset. Scores are generally similar between the two input representations, with a slight numeric advantage in BLEU for the slot-value representation. While the LLMs perform worse on BLEU compared to our seq2seq model, this is expected as they are being used in a zero-shot setting and they are not fine-tuned for data-to-text generation. BERTScores are similar across the 3 models.

For the WebNLG dataset we turn to Table 3. Scores are very similar between slot-value and RDF representations once again, with a slight numeric advantage for the RDF format this time. On this dataset the seq2seq model struggles substantially, with much lower BLEU and BERTScore results compared to the two LLMs, despite the zero-shot

[2] https://unify.ai/; cost breakdown in appendix

[3] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.1

[4] roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.41.1)-rescaled

usage of the LLMs. As this dataset has a much richer semantic space and covers a variety of different topics, data sparsity becomes more of an issue for the seq2seq models, while the LLMs benefit from their very large training data.

	seq2seq		GPT-4o		Llama 3	
	SV	RDF	SV	RDF	SV	RDF
BLEU	47.4	46.9	41.6	39.8	35.8	35.2
BS-F1	0.66	0.66	0.68	0.67	0.63	0.64

Table 2: BLEU and BERTScore F1 results on E2E.

	seq2seq		GPT-4o		Llama 3	
	SV	RDF	SV	RDF	SV	RDF
BLEU	30.2	30.3	47.0	47.8	45.2	45.7
BS-F1	0.35	0.35	0.62	0.64	0.61	0.61

Table 3: Automated evaluation results on WebNLG.

5 Human Evaluation

We asked participants to assess *fluency* and *semantic fidelity*. For fluency, we adapted the questions used by WebNLG 2023 (Cripwell et al., 2023), asking participants to "rate the Output in terms of Fluency" and explaining that "[h]ighly fluent text 'flows well' and is well-connected and free from disfluencies". Participants rated fluency on a 7-point Likert scale ranging from Very Disfluent to Very Fluent. For semantic fidelity (i.e. the faithfulness of the outputs to the inputs), participants saw a table of subjects, predicates, and objects meant to be present in the Output and had to click a radio button to indicate whether that element of the meaning was present, missing, or incorrect. Participants could also indicate if the Output included additional content not present in the Input and had a free text area to describe the inserted content.

For each dataset, we selected 48 inputs from the test across the 7 experimental conditions: the reference text for a control condition plus one text from each system for each input representation. Each participant saw 28 items plus 2 attention check questions presented in a randomised order.

We recruited 36 participants for each dataset through Prolific[5]. We screened participants, requiring them to be first-language speakers of English and resident in a country where English is a majority language (i.e. Australia, Canada, Ireland, New Zealand, South Africa, the United Kingdom, or the United States). The 72 participants completed the task in about 37 minutes (median) and received £7.50 compensation each. The mean participant age was 34 (s.d. 12), with 34 males and 28 females. Our institution approved the study's ethics.

[5] https://www.prolific.com

E2E					
Sys	In	Fluency (sd)	•	○	×
GPT-4o	RDF	6.07 (0.60)	0.95	0.05	0.00
GPT-4o	SV	6.07 (0.74)	0.96	0.04	0.00
Llama 3	RDF	5.94 (1.01)	0.97	0.03	0.00
Llama 3	SV	6.09 (0.70)	0.95	0.05	0.01
s2s	RDF	5.75 (0.90)	0.91	0.07	0.02
s2s	SV	5.74 (0.93)	0.90	0.08	0.02
ref	–	2.80 (1.52)	0.48	0.49	0.03
WebNLG					
Sys	In	Fluency (sd)	•	○	×
GPT-4o	RDF	6.32 (0.82)	0.93	0.04	0.02
GPT-4o	SV	6.33 (0.70)	0.93	0.06	0.01
Llama 3	RDF	6.02 (1.10)	0.89	0.07	0.03
Llama 3	SV	6.18 (1.02)	0.88	0.10	0.02
s2s	RDF	4.12 (1.91)	0.57	0.36	0.08
s2s	SV	4.43 (1.85)	0.54	0.36	0.09
ref	–	5.83 (1.22)	0.93	0.04	0.03

Table 4: Human evaluation results for the E2E Challenge Dataset and the WebNLG Challenge Dataset. Fluency is the mean score on a 7-point Likert scale with standard deviation in parentheses, • is the proportion of inputs expressed correctly, ○ is the proportion which are missing, and × is the proportion which are expressed incorrectly.

5.1 Results & Discussion

Table 4 shows the results, treating fluency ratings ranging from 1-7, where 7 is 'Very Fluent', and reporting the mean and standard deviation. The remaining columns report the proportion of the Input which was Present (•), Missing (○), or Incorrect (×). Both tables show differences between input representations which are much smaller than the standard deviation for each system, though we do observe some differences between the systems. GPT-4o and Llama 3 perform similarly on the E2E corpus, with seq2seq models marginally lower.[6] For WebNLG, the difference in fluency scores for the input representations is larger, though still very small, and the gap between GPT-4o and Llama 3

[6] Scores for reference texts are low for the E2E dataset due to a data preparation error; however, the comparisons between the systems and input types remain valid.

is more pronounced. Here seq2seq performance is worse, with scores lower than the reference texts.

To assess statistical significance, we use an ordinal mixed effects model for the fluency ratings following Howcroft and Rieser (2021), with fixed effects of system and input representation and by-participant random intercepts. The results showed no significant differences for input representation in either dataset. For E2E, there was no significant difference between GPT-4o and Llama 3, though the seq2seq models were significantly worse than both. For WebNLG, both Llama 3 and the seq2seq models performed significantly worse than GPT-4o.

6 Qualitative Error Analysis

Since both LLMs performed well regardless of input representation, we manually examine those instances where they performed worst to see if there are any qualitative patterns.

The two lowest rated GPT-4o texts were scored *Somewhat Disfluent* and both contained the phrase 'located riverside", describing the location of a restaurant. Only one text received a neutral score, and none of these texts had semantic fidelity errors. Three Llama 3 texts scored *Disfluent*, six as *Somewhat Disfluent*, and two as neutral. Llama 3 exhibits a greater tendency to reuse phrases from the input representation in ways that disrupts fluency (e.g. expressing the predicate-object pair `eat type, pub` with the awkward phrase 'is a type of eatery found in a pub'). Sometimes restaurant names are treated as a different kind of entity: 'The Wrestlers' is the name of a restaurant, but Llama 3 treats this as a group of people, producing 'The Wrestlers eat at a pub' instead of 'The Wrestlers is a pub'. Items with the highest proportion of missing or incorrect semantics according to participants tended to be more accurate than reported.

GPT-4o produces one *Disfluent* text for the WebNLG dataset: 'Antwerp International Airport serves the city of Antwerp. The country of Antwerp is Belgium. In Belgium, the language spoken is German.' There are also three *Somewhat Disfluent* and two neutral texts generated. Llama 3 received a *Very Disfluent* rating for a short sentence that is actually fluent: 'Hip hop music is a derivative of Drum and bass'. However, the sentence may have been rated poorly because it is semantically anomalous, or requires domain specific knowledge. Three texts were marked as *Disfluent* and another nine as *Somewhat Disfluent*, some of these seemingly due to awkward phrasing ('Aleksey Chirikov, an icebreaker built in Helsinki, Finland, is led by Juha Sipilä'), and others for being nonsensical, such as 'Atlanta, a city in the United States, is the capital of a country with an ethnic group of Asian Americans, with Washington, D.C. as its capital'. Semantic errors were again infrequent for GPT-4o, though there were more interesting errors for Llama 3. For example, Llama 3 sometimes omits large portions of the meaning representation, expressing only one out of five given predicates.

7 Discussion & Conclusions

We expected that the meaning representation used to encode inputs for neural data-to-text generation would substantially impact either the fluency or the accuracy of generated texts. However, our findings do not support this hypothesis. Instead, we find a strong performance by recent LLMs regardless of input representation, and we find that simpler seq2seq models are also not substantially impacted by these differences. We also observed remarkably few 'hallucinations', or insertions of additional content not present in the input, across both LLMs. We suspect that these results are in part influenced by the fact that both of our source datasets are publicly available and are likely to be included in the training data for both GPT-4o and Llama 3 systems. In future work, we plan to investigate this possibility with the creation of novel, unseen datasets and new linearisations of meaning representations.

8 Limitations & Ethical Considerations

This work explores only two simple meaning representations used for data-to-text generation. For the LLMs, it is possible that they have already seen the data used for our experiments during training.

As mentioned above, our human experiments received institutional ethics oversight.

Acknowledgements

This work was supported by EPSRC project 'NLG for low-resource domains' (EP/T024917/1).

References

AI@Meta. 2024. Llama 3 model card.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the

WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Thiago Castro Ferreira, Helena Vaz, Brian Davis, and Adriana Pagano. 2021. Enriching the E2E dataset. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 177–183, Aberdeen, Scotland, UK. Association for Computational Linguistics.

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

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156.

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

David M. Howcroft and Dimitra Gkatzia. 2023. enunlg: a python library for reproducible neural data-to-text experimentation. In *Proceedings of the 16th International Natural Language Generation Conference: System Demonstrations*, pages 4–5, Prague, Czechia. Association for Computational Linguistics.

David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more underpowered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-

der, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*. ArXiv: 1904.09675.

A Prompting Costs

Llama 3 cost $0.9/1M tokens for both output and input. GPT-4o was $5/1M for input and $15/1M for output. The detailed cost breakdown is provided in Table 5:

	GPT-4o		Llama 3		
	in	out	in	out	TOTAL
e2e_rdf	1.57	1.89	0.28	0.13	3.85
e2e_slot	1.15	1.67	0.21	0.10	3.11
webnlg_rdf	1.10	1.53	0.20	0.09	2.92
webnlg_slot	0.96	1.41	0.17	0.08	2.62
TOTAL	4.78	6.50	0.86	0.40	–

Table 5: Cost in USD for each dataset $\times$ representation with each LLM. Totals do not sum up exactly across the margins due to rounding to the nearest penny.

B Example Outputs

Here we show a few example texts from our seq2seq model as well as GPT-4o and Llama 3.

E2E
<SUBJECT> Blue Spice <PREDICATE> eat type <OBJECT> coffee shop <TRIPLE_SEP> <SUBJECT> Blue Spice <PREDICATE> area <OBJECT> city centre Blue Spice is a pub located in the city centre.
name == Blue Spice <PAIR_SEP> eat type == coffee shop <PAIR_SEP> area == city centre Blue Spice is a pub in the city centre.
<SUBJECT> Clowns <PREDICATE> eat type <OBJECT> coffee shop <TRIPLE_SEP> <SUBJECT> Clowns <PREDICATE> customer rating <OBJECT> 5 out of 5 <TRIPLE_SEP> <SUBJECT> Clowns <PREDICATE> near <OBJECT> Crowne Plaza Hotel Clowns is a pub near Crowne Plaza Hotel with a customer rating of 5 out of 5.
name == Clowns <PAIR_SEP> eat type == coffee shop <PAIR_SEP> customer rating == 5 out of 5 <PAIR_SEP> near == Crowne Plaza Hotel Clowns is a pub near Crowne Plaza Hotel with a customer rating of 5 out of 5.
WebNLG
<SUBJECT> Adolfo Suárez Madrid– Barajas Airport <PREDICATE> location <OBJECT> Madrid , Paracuellos de Jarama , San Sebastián de los Reyes and Alcobendas Adolfo Suárez Madrid–Barajas Airport is located in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes and Alcobendas.
name == Adolfo Suárez Madrid– Barajas Airport <PAIR_SEP> location == Madrid , Paracuellos de Jarama , San Sebastián de los Reyes and Alcobendas Adolfo Suárez Madrid–Barajas Airport is located in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes and Alcobendas.
<SUBJECT> William Anders <PREDICATE> date of retirement <OBJECT> 1969-09-01 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> was selected by nasa <OBJECT> 1963 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> time in space <OBJECT> 8820.0"(minutes) <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> birth date <OBJECT> 1933-10-17 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> occupation <OBJECT> Fighter pilot <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> birth place <OBJECT> British Hong Kong <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> was a crew member of <OBJECT> Apollo 8 William Anders was born in __PATIENT-5__ and graduated from UT Austin in 1955 with a B. S. He served as a test pilot and was selected by NASA in Apollo 8. He was selected by NASA in __PATIENT-4__.
name == William Anders <PAIR_SEP> date of retirement == 1969-09-01 <PAIR_SEP> was selected by nasa == 1963 <PAIR_SEP> time in space == 8820.0"(minutes) <PAIR_SEP> birth date == 1933-10-17 <PAIR_SEP> occupation == Fighter pilot <PAIR_SEP> birth place == British Hong Kong <PAIR_SEP> was a crew member of == Apollo 8 William Anders was born on March 15, 1932 in He graduated from NWC with an M. A. in 1957. He was a test pilot and died in St Louis.

Table 6: Example input-output pairs from both datasets and input representations with our seq2seq model.

E2E
<SUBJECT> Blue Spice <PREDICATE> eat type <OBJECT> coffee shop <TRIPLE_SEP> <SUBJECT> Blue Spice <PREDICATE> area <OBJECT> city centre Blue Spice is a coffee shop located in the city centre.
name == Blue Spice <PAIR_SEP> eat type == coffee shop <PAIR_SEP> area == city centre Blue Spice is a coffee shop located in the city centre.
<SUBJECT> Clowns <PREDICATE> eat type <OBJECT> coffee shop <TRIPLE_SEP> <SUBJECT> Clowns <PREDICATE> customer rating <OBJECT> 5 out of 5 <TRIPLE_SEP> <SUBJECT> Clowns <PREDICATE> near <OBJECT> Crowne Plaza Hotel Clowns enjoy eating at a coffee shop. They have a customer rating of 5 out of 5. They are located near the Crowne Plaza Hotel.
name == Clowns <PAIR_SEP> eat type == coffee shop <PAIR_SEP> customer rating == 5 out of 5 <PAIR_SEP> near == Crowne Plaza Hotel Clowns is a coffee shop near the Crowne Plaza Hotel with a customer rating of 5 out of 5.
WebNLG
<SUBJECT> Adolfo Suárez Madrid– Barajas Airport <PREDICATE> location <OBJECT> Madrid , Paracuellos de Jarama , San Sebastián de los Reyes and Alcobendas Adolfo Suárez Madrid–Barajas Airport is located in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes, and Alcobendas.
name == Adolfo Suárez Madrid– Barajas Airport <PAIR_SEP> location == Madrid , Paracuellos de Jarama , San Sebastián de los Reyes and Alcobendas Adolfo Suárez Madrid–Barajas Airport is located in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes, and Alcobendas.
<SUBJECT> William Anders <PREDICATE> date of retirement <OBJECT> 1969-09-01 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> was selected by nasa <OBJECT> 1963 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> time in space <OBJECT> 8820.0"(minutes) <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> birth date <OBJECT> 1933-10-17 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> occupation <OBJECT> Fighter pilot <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> birth place <OBJECT> British Hong Kong <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> was a crew member of <OBJECT> Apollo 8 William Anders, born on October 17, 1933, in British Hong Kong, was selected by NASA in 1963. He served as a fighter pilot and was a crew member of Apollo 8. He spent 8,820 minutes in space and retired on September 1, 1969.
name == William Anders <PAIR_SEP> date of retirement == 1969-09-01 <PAIR_SEP> was selected by nasa == 1963 <PAIR_SEP> time in space == 8820.0"(minutes) <PAIR_SEP> birth date == 1933-10-17 <PAIR_SEP> occupation == Fighter pilot <PAIR_SEP> birth place == British Hong Kong <PAIR_SEP> was a crew member of == Apollo 8 William Anders, born on October 17, 1933, in British Hong Kong, was a fighter pilot selected by NASA in 1963. He spent 8,820 minutes in space and was a crew member of Apollo 8. Anders retired on September 1, 1969.

Table 7: Example input-output pairs from both datasets and input representations with GPT-4o.

E2E
<SUBJECT> Blue Spice <PREDICATE> eat type <OBJECT> coffee shop <TRIPLE_SEP> <SUBJECT> Blue Spice <PREDICATE> area <OBJECT> city centre Blue Spice is a type of coffee shop located in the city centre.
name == Blue Spice <PAIR_SEP> eat type == coffee shop <PAIR_SEP> area == city centre Blue Spice is a coffee shop in the city centre.
<SUBJECT> Clowns <PREDICATE> eat type <OBJECT> coffee shop <TRIPLE_SEP> <SUBJECT> Clowns <PREDICATE> customer rating <OBJECT> 5 out of 5 <TRIPLE_SEP> <SUBJECT> Clowns <PREDICATE> near <OBJECT> Crowne Plaza Hotel Clowns eat at a coffee shop, have a 5 out of 5 customer rating, and are located near the Crowne Plaza Hotel.
name == Clowns <PAIR_SEP> eat type == coffee shop <PAIR_SEP> customer rating == 5 out of 5 <PAIR_SEP> near == Crowne Plaza Hotel Clowns is a 5-star coffee shop near the Crowne Plaza Hotel.
WebNLG
<SUBJECT> Adolfo Suárez Madrid– Barajas Airport <PREDICATE> location <OBJECT> Madrid , Paracuellos de Jarama , San Sebastián de los Reyes and Alcobendas Adolfo Suárez Madrid–Barajas Airport is located in Madrid, as well as the surrounding areas of Paracuellos de Jarama, San Sebastián de los Reyes, and Alcobendas.
name == Adolfo Suárez Madrid– Barajas Airport <PAIR_SEP> location == Madrid , Paracuellos de Jarama , San Sebastián de los Reyes and Alcobendas Adolfo Suárez Madrid–Barajas Airport is located in Madrid, Paracuellos de Jarama, San Sebastián de los Reyes, and Alcobendas.
<SUBJECT> William Anders <PREDICATE> date of retirement <OBJECT> 1969-09-01 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> was selected by nasa <OBJECT> 1963 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> time in space <OBJECT> 8820.0"(minutes) <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> birth date <OBJECT> 1933-10-17 <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> occupation <OBJECT> Fighter pilot <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> birth place <OBJECT> British Hong Kong <TRIPLE_SEP> <SUBJECT> William Anders <PREDICATE> was a crew member of <OBJECT> Apollo 8 William Anders retired on September 1, 1969. He was selected by NASA in 1963. Anders spent 8820 minutes in space. He was born on October 17, 1933, in British Hong Kong. He was a fighter pilot and a crew member of Apollo 8.
name == William Anders <PAIR_SEP> date of retirement == 1969-09-01 <PAIR_SEP> was selected by nasa == 1963 <PAIR_SEP> time in space == 8820.0"(minutes) <PAIR_SEP> birth date == 1933-10-17 <PAIR_SEP> occupation == Fighter pilot <PAIR_SEP> birth place == British Hong Kong <PAIR_SEP> was a crew member of == Apollo 8 William Anders, a fighter pilot born on October 17, 1933, in British Hong Kong, was selected by NASA in 1963. He was a crew member of Apollo 8 and spent 8820 minutes in space before retiring on September 1, 1969.

Table 8: Example input-output pairs from both datasets and input representations with Llama 3.