MURMUR: Modular Multi-Step Reasoning for Semi-Structured Data-to-Text Generation

Swarnadeep Saha[1] **Xinyan Velocity Yu**[2] **Mohit Bansal**[1]
Ramakanth Pasunuru[2] **Asli Celikyilmaz**[2]
[1]UNC Chapel Hill [2]Meta AI
{swarna, mbansal}@cs.unc.edu
{velocityyu, rpasunuru, aslic}@meta.com

Abstract

Prompting large language models has enabled significant recent progress in multi-step reasoning over text. However, when applied to text generation from semi-structured data (e.g., graphs or tables), these methods typically suffer from low semantic coverage, hallucination, and logical inconsistency. We propose MURMUR, a neuro-symbolic modular approach to *text generation from semi-structured data with multi-step reasoning*. MURMUR is a best-first search method that generates reasoning paths using: (1) neural and symbolic modules with specific linguistic and logical skills, (2) a grammar whose production rules define valid compositions of modules, and (3) value functions that assess the quality of each reasoning step. We conduct experiments on two diverse data-to-text generation tasks like WebNLG and LogicNLG. The tasks differ in their data representations (graphs and tables) and span multiple linguistic and logical skills. MURMUR obtains significant improvements over recent few-shot baselines like direct prompting and chain-of-thought prompting, while also achieving comparable performance to fine-tuned GPT-2 on out-of-domain data. Moreover, human evaluation shows that MURMUR generates highly faithful and correct reasoning paths that lead to 26% more logically consistent summaries on LogicNLG, compared to direct prompting.[1]

1 Introduction

Data-to-text generation (McKeown, 1992; Reiter and Dale, 1997; Wen et al., 2015; Dušek and Jurcicek, 2015; Mei et al., 2016; Novikova et al., 2017; Gatt and Krahmer, 2018) is the task of generating fluent, faithful, and consistent summaries of semi-structured data. Recent works have introduced different data-to-text generation tasks wherein the data is represented in diverse structures, like meaning representations (Novikova et al., 2017), graphs (Gardent et al., 2017), or tables (Lebret et al., 2016; Parikh et al., 2020; Chen et al., 2020a). Text generation from such data is challenging because it extends surface realization of the input content and requires various reasoning and compositionality skills, such as filtering a table based on a certain criterion, retrieving the maximum value from a table column, etc.

Existing works fine-tune pre-trained language models (Radford et al., 2019; Raffel et al., 2020) as the de-facto standard for building supervised

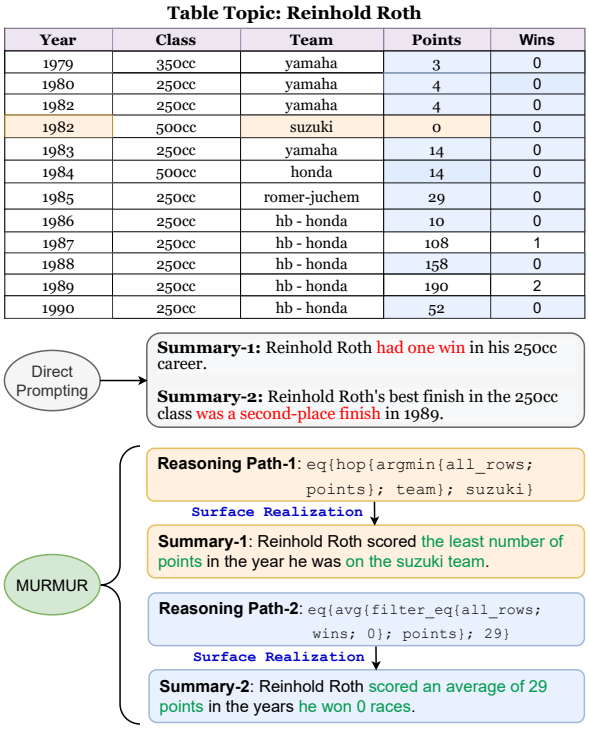

Figure 1: Sample table from LogicNLG and two logical summaries generated by MURMUR and Direct Prompting baseline. Direct Prompting summaries include logical inconsistencies and hallucinations (marked in red) while MURMUR generates reasoning paths (composed of modules) and converts them to logically consistent summaries (marked in green). Each color code highlights part of the table relevant to a MURMUR summary.

[1]Supporting code available at https://github.com/swarnaHub/MURMUR

Findings of the Association for Computational Linguistics: ACL 2023, pages 11069–11090
July 9-14, 2023

data-to-text generation systems (Kale and Rastogi, 2020; Agarwal et al., 2021). However, this requires a large amount of domain-specific parallel data, which is expensive to obtain, and training models on such data also affects out-of-domain generalization (Laha et al., 2020; Dušek et al., 2020).

Motivated by the recent success of few-shot prompting in multi-step reasoning over text (Wei et al., 2022; Nye et al., 2021; Wang et al., 2022a; Dohan et al., 2022), we pose data-to-text generation as *multi-step reasoning over data*.[2] Reasoning over data for text generation brings its own set of challenges: (1) **Generation Quality**: Firstly, directly prompting large language models (LLMs) can cause models to suffer from low semantic coverage, hallucinations, and logically inconsistent generations (see red marked phrases for the Direct Prompting summaries in Fig. 1). Other prompting methods like Chain-of-Thought (CoT) encourage LLMs to also generate intermediate reasoning steps (Wei et al., 2022) but it compromises the *transparency*, *faithfulness*,[3] and *correctness* of the reasoning process due to the lack of explicit conditioning between the reasoning steps (Creswell and Shanahan, 2022). (2) **Transformation-invariance:** Text is a sequence of tokens while data is typically represented as a *set* of elements (e.g., a graph is a set of edges, a table is a set of rows, etc). Hence, a model that reasons over data must be *transformation-invariant* (Wang et al., 2022a). For instance, the summary generated from a table should be invariant to randomly shuffling the rows of the table. Thus, prompting methods that linearize the data in an arbitrary order, can be prone to some variance (see Table 3 and 6).

We propose MURMUR, a few-shot **Mod**u**l**a**r** **Mu**lti-step **R**easoning approach to text generation from data (§3). It is a best-first search algorithm (§3.4) that generates reasoning paths (see examples in Fig 1) with three features: (1) **Modularity** (§3.1): MURMUR defines a set of few-shot neural and symbolic modules with diverse input/output data types that constitute multiple steps in a reasoning path. Neural modules perform linguistic skills that LLMs are good at (e.g., the *Surface Realization* module in Fig. 1 converts a reasoning path to a natural language summary) and symbolic modules perform logical skills that they mostly struggle with (Wang et al., 2022b; Gao et al., 2022) (e.g., the *argmin* module in Fig. 1 finds the row with the minimum points); (2) **Grammar** (§3.2): MURMUR introduces a grammar whose production rules specify valid compositions of modules. For instance, in the second path of Fig. 1, MURMUR first generates the module *filter_eq* followed by the *avg* module, because the former outputs a *table* data type which is also the input data type to the latter; (3) **Value functions** (§3.3): To evaluate the quality of each plausible reasoning step and choose the best modules at each step, MURMUR defines value functions that score, rank, and select the best steps. For example, in the second path of Fig. 1, an *avg* module is perhaps more salient than a *max* or *min* module (which only finds the maximum or minimum points).

Our **findings** are: MURMUR can perform multi-step generative reasoning on simple to complex semi-structured data-to-text generation tasks including WebNLG (Gardent et al., 2017), a graph-to-text task (§5) and LogicNLG (Chen et al., 2020a), a table-to-text task (§6). We compare MURMUR with state-of-the-art supervised (end-to-end and pipeline) and few-shot prompting methods. On WebNLG, MURMUR obtains significant improvements in semantic coverage and hallucinations of generated summaries over other few-shot baselines like direct prompting and CoT prompting. Additionally, MURMUR demonstrates good out-of-domain generalizability by obtaining comparable performance to fine-tuned LMs like GPT-2. On LogicNLG, human evaluation demonstrates that MURMUR significantly improves the logical consistency of summaries over direct prompting (by up to 26%), showcasing the strength of a neuro-symbolic approach for data-to-text generation.

2 Definitions: Reasoning Step and Path

A *Reasoning Step* is a triple $(\mathcal{M}, \mathcal{X}, y)$ where a module $\mathcal{M}$ performs a certain skill by conditioning on an input $\mathcal{X}$ to generate an output y. For example, in Fig. 2, the module *argmin* takes a table and a column (*points*) as input and outputs the row with the minimum points. A *Reasoning Path* is defined as a sequence of such reasoning steps $\{(\mathcal{M}_i, \mathcal{X}_i, y_i)\}_{i=1}^{r}$. Fig. 2 shows an example of a reasoning path, represented as a nested structure. It consists of three reasoning steps for three modules (*argmin*, *hop*, and *eq*). The *argmin* module outputs

[2] By data, we mean semi-structured data such as graphs or tables. Henceforth, we will refer to it as just 'data'.

[3] Faithful reasoning refers to an underlying causal structure in the reasoning process. This is different from a text's faithfulness to an input context which will be called hallucinations.

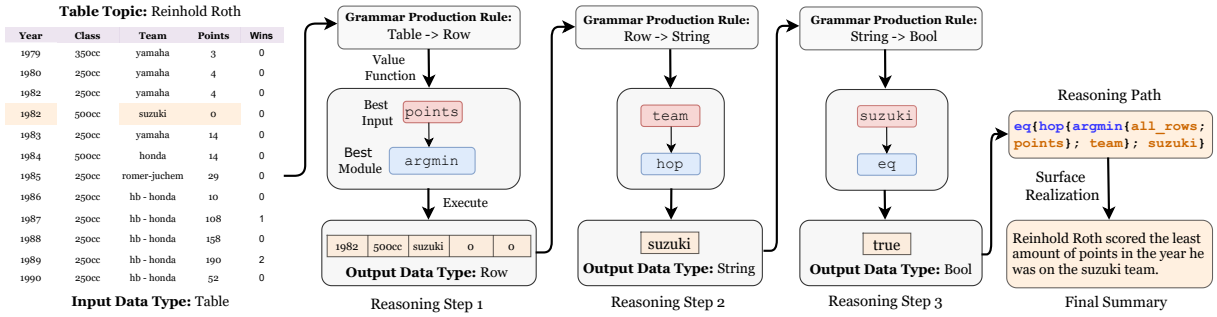

Figure 2: Illustration of MURMUR generating a reasoning path and then converting it into a logically consistent summary, supported by the input table. The reasoning path consists of three reasoning steps. At each step, MURMUR chooses a set of plausible modules (according to a grammar) and then selects the best module, with the best input according to a value function. The output generated at each step serves as the input to the next step.

the row in the table with minimum points, which is the input to the next module *hop* that selects a column from that row. MURMUR generates textual summaries by constructing such reasoning paths that are then converted to the final outputs through a *Surface Realization* module, as shown in Fig. 2.[4]

3 MURMUR Approach

MURMUR consists of four components: (1) a set of modules, (2) a grammar, (3) value function(s), and (4) a search algorithm that brings all the previous three components together. The search algorithm constructs reasoning paths by first identifying plausible modules at each reasoning step according to the grammar and then determining the best modules (and their corresponding inputs) with the help of value functions. Fig. 2 shows a working example of MURMUR, in which given an input table, it searches for a reasoning path (of three steps), and finally converts it into a summary. The specifics of MURMUR's components vary based on the task at hand. As case studies, we consider two data-to-text generation tasks: WebNLG (Gardent et al., 2017), a graph-to-text generation task and LogicNLG (Chen et al., 2020a), a complex table-to-text generation task where the goal is to generate logical summaries from salient parts of the table.

3.1 MURMUR Modules

MURMUR defines a set of modules $\{\mathcal{M}_i\}_{i=1}^{m}$ that perform specialized reasoning skills for the corresponding task. Formally, each module $\mathcal{M}_i$ is defined as a multi-variate function $\mathcal{M}_i : \mathcal{X} \rightarrow y$ that maps an n-tuple input $\mathcal{X} = (x_1, \cdots, x_n)$ to an output y. Each input variable x_i and output y can have their own expected data types d_i and d_y respectively. These data types could be user-defined[5] like *Table*, *Triple*, etc or standard ones like *String*, *Number*, *Bool*, etc. For example, in Fig. 2, the module $\mathcal{M}_{argmin} : (t, c) \rightarrow r$ takes a table t (with data type *Table)* and a column c (with data type *String*) as input and outputs a row r (with data type *row*) with the minimum value in column c. The modules are implemented as few-shot neural models or symbolic functions. We choose few-shot neural modules for linguistic skills that LLMs typically excel at and symbolic modules for logical operations that LLMs mostly struggle with (Wang et al., 2022b; Gao et al., 2022). Reasoning over semi-structured data allows us to implement symbolic modules with PYTHON functions. Below we provide examples of neural and symbolic modules for the two tasks.

Neural Linguistic Modules. In any modular data-to-text generation approach, one of the modules is responsible for the transition from structured data to unstructured text. We call it Surface Realization. In particular, for WebNLG, we define it as $\mathcal{M}_{sr} : t \rightarrow s$ that converts a triple t (with data type *Triple*) into a short sentence s (with data type *String*). For LogicNLG, we define it as $\mathcal{M}_{sr} : (t, p) \rightarrow s$ that takes a table t (with data type *Table*) and a reasoning path p as input and converts it into a summary s (with data type *String*). As we show later, in WebNLG, *Surface Realization*s are the first reasoning steps, while in LogicNLG, it

[4]For better illustration, we show Surface Realization outside of the reasoning path but ideally, it can be considered as another (final) step in the reasoning path.

[5]The modules are analogous to function definitions with expected IO types. Similarly, user-defined data types can be thought of as class definitions. For instance, a data type *Triple* can be implemented as a class consisting of a subject, a relation, and an object (all with data type *String*).

Module	Description
Filtering	$\mathcal{M}_{filter} : (t, cr) \rightarrow t'$: takes a table t and a filtering criterion cr as input and outputs a table t' with rows where the criterion cr is satisfied.
Aggregation	Performs aggregation operations on a table. For example, $\mathcal{M}_{max} : (t, c) \rightarrow n$ is a *max* module that takes a table t and a column c as input and outputs the maximum number n in column c.
Boolean	$\mathcal{M}_{bool} : (t, cr) \rightarrow b$: takes a table t and a criterion cr as input and outputs a boolean b based on whether the criterion is satisfied.
Hop	$\mathcal{M}_{hop} : (r, c) \rightarrow e$: takes a row r and a column c as input; outputs the element e in (r, c) cell.

Table 1: MURMUR Symbolic Modules to perform logical operations over tables in Table-to-Text generation.

is the last step. For WebNLG, we also define a *Text Fusion* module $\mathcal{M}_{tf} : (s_1, s_2) \rightarrow s$ that combines two strings s_1 and s_2 into a coherent text s. *Text Fusion* iteratively combines intermediate generations at each step, enabling more controllability in generation (Tan et al., 2021).

Symbolic Logical Modules. For LogicNLG, drawing motivation from prior work (Chen et al., 2020c), we define different categories of symbolic modules that perform logical operations over tables (see Table 1 and refer to Table 8 for the detailed list). WebNLG requires summarizing an input graph and hence, does not involve any logical modules.

3.2 Grammar over Modules

The role of the grammar is to determine a set of plausible modules in a reasoning step and how they should be composed. The production rules of the grammar capture possible transitions from an input data type to an output data type(s) (see Fig. 2 and Table 2). Each production rule thus defines multiple permissible modules. For example, the production rule 'Table → Number' (meaning that a number can be generated from a table) is valid for both *max* and *min* modules. When MURMUR searches for reasoning paths, the grammar reduces the search space (over all possible modules) by only selecting the ones that can be composed at each reasoning step. We provide examples below of how such grammars are constructed.

Grammar for Graph-to-Text Generation. Table 2 shows the grammar for Graph-to-Text generation. It consists of two production rules, one for Surface Realization and another for Text Fusion. Past pipeline approaches for graph-to-text generation (Xiang et al., 2022) also perform surface realization followed by fusion, as explained through the grammar.

Graph-to-Text (WebNLG)
Triple → String (String, String) → String
Table-to-Text (LogicNLG)
Table → Table | Row | Number | Boolean Row → String | Number String | Number → Boolean (Table, Path) → String

Table 2: Grammars for WebNLG and LogicNLG defining production rules between different data types.

Grammar for Table-to-Text Generation. Generating logical summaries from a table is a more challenging task. Based on the types of modules introduced previously, we define a grammar, as shown in Table 2. As an instance, the first rule encodes the knowledge that given an input of type *Table*, one can output a *Table*, a *Row* of the table, a *Number*, or a *Boolean*.

3.3 Value Functions

While the grammar helps reduce the search space by defining permissible compositions of modules, each reasoning step can still have multiple plausible modules and each module can also have multiple plausible inputs to choose from. Thus, MURMUR introduces value functions (see Fig. 2) that assess the quality of each plausible reasoning step by scoring, ranking, and selecting the best step(s).

Value Function for Graph-to-Text Generation. In a Graph-to-Text generation task, each intermediate reasoning step r generates a summary y_r for a subset of edges (triples) G_r from the input graph (see Fig. 7 for an illustration). The value functions evaluate the following two aspects of the generated summary y_r. First, **Fluency** is measured by log-likelihood of the generated text similar to BARTScore (Yuan et al., 2021):

$$S_f(y_r) = \exp\{\frac{1}{l}\sum_{i=1}^{l} \log p_\theta(y_r^i | y_r^{<i})\}$$

Second, **Semantic Consistency** measures the average logical entailment probability $P_e(\cdot)$ between the generation y_r and triples G_r[6] and vice-versa:

$$S_{sc}(G_r, y_r) = 0.5 \times (P_e(G_r, y_r) + P_e(y_r, G_r))$$

We use an NLI model to compute entailment probabilities. The both-way entailment scores capture

[6]We concatenate the surface realizations of the triples to construct the sequence for the NLI model.

equivalence between the triples and the generation, ensuring that the latter not only covers all the triples but also does not hallucinate any new information. Overall score is an ensemble of the two scores, given by $\alpha S_f(y_r) + (1-\alpha)S_{sc}(G_r, y_r)$.

Value Function for Table-to-Text Generation. Our value function chooses the best module(s) at each reasoning step, as well as the best input(s) for the corresponding module(s).[7] For instance, if a reasoning step generates a number from a table (according to the grammar), the value function should determine the best module(s) between *max*, *min*, etc, as well as which column the *max* or *min* module should be computed on. Taking inspiration from past work on verifying intermediate reasoning traces over text (Creswell and Shanahan, 2022; Yang et al., 2022), we train a value function $S : (T, P_r) \rightarrow p$ that judges the correctness of a partial reasoning path P_r for an input table T. In particular, we train a binary classifier on samples with correct and incorrect partial reasoning paths. We call this value function a *saliency metric* because it selects the best reasoning steps that reason over salient parts of the table. We discuss the model and training data for our saliency metric in § 4.2.

3.4 Search Algorithm

We now describe how all the three components discussed above come together in generating reasoning paths for MURMUR (see Fig. 2). We propose a best-first search algorithm that operates as follows. It takes as input a set of m modules $\{\mathcal{M}_i\}_{i=1}^{m}$, a grammar $\mathcal{G}$, a value function $\mathcal{V}$, and number of reasoning paths or summaries to generate p. Additionally, it considers a hyperparameter, the beam size b of the search ($b \geq p$). The search begins by initializing an empty priority queue that maintains the beam (best b partial reasoning paths to be explored anytime during the search). Next, at each step, MURMUR (1) pops an element from the queue, (2) identifies the data type of the element (e.g., *Table*), (3) looks up the grammar to find all possible transitions from that data type (e.g., *Row*, *Number*), (4) selects all modules for each such transition (e.g., *argmax* and *argmin* for 'Table → Row', *max* and *min* for 'Table → Number'), and (5) constructs all plausible reasoning steps consisting of modules and their corresponding inputs (e.g., all numerical columns for *argmax*). It then scores all these reasoning steps using the value function, ranks them, and only keeps the top-b paths in the queue. For WebNLG, the search terminates when all triples have been iterated. For LogicNLG, a reasoning path is complete when the current module outputs a boolean variable that evaluates to true (e.g., the *eq* module). Upon termination of the search, we return the top-p paths and the corresponding summaries.

[7] In Graph-to-Text, we only need to choose the best inputs because at each step there is only one plausible module.

4 Experimental Setup

4.1 Graph-to-Text Generation

We report results on both seen and unseen splits of the test set of WebNLG (Gardent et al., 2017).[8]

Modules. We implement both modules, *Surface Realization* and *Text Fusion* as few-shot neural models by prompting OPT-175B (Zhang et al., 2022) with skill-specific prompts (see Appendix D) and greedy decoding.

Value Function. As defined in §3.3, we compute fluency using the log probabilities estimated by OPT-175B. The entailment probability for the semantic scorer is based on a DeBERTa-base model (He et al., 2020) trained on a collection of eight NLI datasets.[9] The mixing ratio α is set to 0.05. At each reasoning step, MURMUR scores and ranks the intermediate generations in the queue using the value function. Subsequently, it only explores the highest scoring intermediate generation in the next step of the search and prunes the rest.[10]

4.2 Table-to-Text Generation

Modules. We implement all logical modules, as described in §3.1, with PYTHON functions. We again prompt OPT-175B for the *Surface Realization* module (see Appendix D for the prompt).

Value Function. Our saliency metric is a binary classifier. Specifically, we train a BERT-base model that takes a table (as a sequence of rows) and a partial reasoning path as input and classifies it as correct or incorrect. During inference, we consider the correct class probability as the saliency score. We obtain training data from the Logic2Text dataset (Chen et al., 2020c) that annotates open-domain tables with gold reasoning paths. Given a

[8] The 2017 version of the dataset is available at https://webnlg-challenge.loria.fr/challenge_2017/.

[9] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c.

[10] This is equivalent to performing greedy search. We experimented with larger beams for a more exhaustive search without observing any noticeable improvement in performance.

		BLEU			METEOR		
		Seen	Unseen	All	Seen	Unseen	All
supervised	MELBOURNE†	54.5	33.2	45.1	41.0	33.0	37.0
	GPT-2-large†	65.3	43.1	55.5	46.0	38.0	42.0
	T5-large†	64.9	54.0	59.9	46.0	43.0	44.0
	Neural Pipeline‡	-	-	43.3	-	-	39.3
few-shot	Direct Prompting (k=1)⋆	33.1±0.3	34.2±0.1	33.6±0.1	30.4±0.1	31.2±0.1	30.8±0.1
	Direct Prompting (k=5)⋆	39.9±0.3	38.9±0.3	39.5±0.1	34.3±0.1	34.3±0.3	34.4±0.1
	CoT Prompting (k=1)⋆	22.2±0.2	14.9±0.2	18.0±0.1	22.3±0.1	22.9±0.2	22.6±0.1
	MURMUR (k=1)⋆	**41.4±0.0**	**41.1±0.0**	**41.3±0.0**	**37.1±0.0**	**37.1±0.0**	**37.1±0.0**

Table 3: Comparison of supervised and few-shot approaches on the WebNLG Seen and Unseen splits of the test set. † = Supervised with 7k in-domain samples. ‡ = Supervised with a synthetic corpus of 934k samples. ⋆ = Few-shot with k demonstrations. We report mean and variance for all few-shot methods with three random triple orderings.

gold reasoning path, we create *correct* partial paths by breaking it at each intermediate step/module and *incorrect* paths by performing two types of perturbations on every correct partial path: (1) replacing the module at the current step with another module of same data type (e.g., replacing module *max* with module *min*); (2) replacing the inputs to the module with other plausible inputs (e.g., replacing *max* over column c_1 with *max* over another column c_2). See Appendix C.4 for an illustration of the training data creation process. We choose 221 (table, reasoning path) pairs from the Logic2Text dataset and convert them into 1500 correct and incorrect training samples consisting of (table, partial reasoning path) pairs. While choosing the samples, we ensure that the corresponding tables have no overlap with those in the test and validation sets of LogicNLG. We choose the beam size of the search to be 20 (see further analysis of beam sizes in Appendix C.3).

5 Experiments on Graph-to-Text

5.1 Comparison of MURMUR with supervised and few-shot methods

Baselines. We compare with both supervised and few-shot baselines.

- **Supervised.** We compare with *MELBOURNE*, a non-pretrained encoder-decoder model (Gardent et al., 2017) and two fine-tuned LMs, *GPT-2-large* (Radford et al., 2019) and *T5-large* (Raffel et al., 2020). We also compare with a SOTA modular pipeline approach, *Neural Pipeline* (Kasner and Dušek, 2022) that first converts triples to sentences using hand-designed templates and subsequently orders and fuses the sentences by fine-tuning on a large synthetic corpus of 934k samples.
- **Few-shot.** For direct comparisons, we consider two few-shot baselines, *Direct Prompting (DP)* that directly prompts the OPT-175B model to generate a summary of the graph, and *Chain-of-Thought Prompting (CoT)* (Wei et al., 2022) that prompts the model to generate the summary step-by-step (see Appendix D for prompts). We choose the demonstrations randomly from the training data and keep them consistent across all few-shot methods.

Metrics. Following prior work, we perform automatic evaluation using BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).

Results. For all few-shot methods, we report mean and variance of three random triple orderings. Table 3 shows the results. MURMUR significantly outperforms DP and CoT by up to 8 points in BLEU and METEOR ($p < 0.001$), when using a single demonstration (k=1).[11] MURMUR even outperforms DP with five demonstrations (k=5). Prompting an LLM by simply concatenating the intermediate steps for CoT does not work well for text generation. MURMUR also outperforms a supervised baseline MELBOURNE and obtains comparable performance to fine-tuned models like GPT-2 on the unseen test split. Through its modular treatment, MURMUR generates outputs with more coverage of triples and lesser hallucinations, as reflected in the improved scores and further demonstrated in §5.2 through human evaluation. Finally, MURMUR is transformation-invariant because it treats the graph as a *set* (not *sequence*) of triples. Refer to Appendix B for experiments studying the number and variation in demonstrations.

[11] A single demonstration for DP can get decomposed into multiple in-context examples in MURMUR. However, like CoT, the decompositions are all from the same demonstration.

	DP	MURMUR	% Improve
Omissions↓	1.64±0.06	**0.73±0.01**	+24%
Hallucinations↓	0.77±0.03	**0.43±0.03**	+9%
Disfluencies↓	**0.14±0.05**	0.30±0.04	-4%

Table 4: Average count of omissions, hallucinations, and disfluencies in WebNLG summaries.

5.2 Human Evaluation of Final Summaries and Intermediate Reasoning Steps

Evaluation of Final Summaries. We compare the summaries generated by DP (our best baseline) and MURMUR. Two NLP experts take part in the study with 50 randomly chosen test samples (having an average of 3.8 triples). They count the number of omissions, hallucinations, and disfluencies in the generated outputs.[12] Our results in Table 4 demonstrate that MURMUR benefits significantly from a step-wise generative process and reduces omissions by 24% and hallucinations by 9%. We do observe a slight drop in fluency in MURMUR's generations because of its iterative fusion process.

Evaluation of Intermediate Reasoning Steps. We also evaluate the quality of the individual reasoning steps of MURMUR. For every reasoning step of a data point, we provide the annotators with the (1) generation, and (2) the previous steps that the current step is conditioned on. We conduct this study on six randomly chosen test examples, spanning 50 reasoning steps (28 Surface Realization and 22 Text Fusion). Annotators judge the generations for their grammaticality, module faithfulness (i.e., if the module is doing what it is supposed to do), and correctness (e.g., whether the fusion is correct). From Table 5, we conclude that both modules are almost always grammatical, and highly faithful and 64% of fusion operations are also fully correct.

6 Experiments on Table-to-Text

6.1 Comparison of MURMUR with supervised and few-shot methods

Baselines. We compare with several non-pretrained and pretrained supervised methods as well as few-shot methods.

- **Non-pretrained Supervised.** We compare MURMUR with a non-pretrained transformer model, *Field-Infusing + Trans* (Chen et al., 2020a).

[12] All metrics including hallucinations and disfluencies are counted at the level of triples or facts. This allows us to compare the raw counts of both methods with respect to the average number of triples in the input samples.

Module	Grammatical	Faithful	Correct
Surface Realization	1.00	1.00	0.82
Text Fusion	0.90	0.72	0.64

Table 5: Fraction of grammatical, module faithful, and correct intermediate reasoning steps generated by the two modules in MURMUR for WebNLG.

	BLEU-1 / BLEU-2 / BLEU-3
Field-Infusing†	43.7 / 20.9 / 8.4
BERT-TabGen†	49.1 / 27.7 / 13.5
GPT-TabGen†	49.6 / 28.2 / 14.2
GPT-Coarse-to-Fine†	49.0 / 28.3 / 14.6
DCVED†	49.5 / 28.6 / 15.3
Direct Prompting⋆	37.2±0.4 / 18.8±0.2 / 8.6±0.2
CoT Prompting⋆	35.6±0.2 / 18.6±0.1 / 8.8±0.0
BART + SR‡	39.2±0.2 / 20.6±0.2 / 9.5±0.0
MURMUR‡	**39.8±0.0 / 22.2±0.0 / 11.2±0.0**
- saliency⋆	39.6±0.0 / 21.9±0.0 / 10.6±0.0

Table 6: Comparison of supervised non-pretrained, pretrained, and few-shot approaches on the LogicNLG test set. † = Supervised with 37k in-domain samples. ⋆ = Few-shot with 1 demonstration. ‡ = Few-shot with 1 demonstration and 221 gold (table, path) pairs. We report mean and variance for all few-shot methods with three random orderings of the input table rows.

- **Pretrained Supervised.** Next, we compare with three pre-trained LMs based on BERT and GPT-2, *BERT-TabGen*, *GPT-TabGen*, *GPT-Coarse-to-Fine* (Chen et al., 2020a) and a deconfounded variational encoder-decoder model, *DCVED* (Chen et al., 2021).
- **Few-shot.** We also compare with *Direct Prompting (DP)* and *CoT Prompting*. Additionally, we evaluate the effect of our search algorithm and saliency metric. First, in *BART + SR*, instead of searching for reasoning paths, we fine-tune a BART model that generates reasoning paths in one go. As training data, we leverage the (table, gold reasoning path) pairs that are used for training the saliency metric. The surface realization (SR) step is left unchanged. Second, we remove the saliency metric by randomly selecting a module at each step (but according to the grammar). All few-shot methods use one random demonstration.

Metrics. Following Chen et al. (2020a), we compare all methods with BLEU scores. They also propose metrics to evaluate logical consistency but we found such learned metrics do not correlate well with humans. Instead, we conduct more reliable human evaluations of logical correctness in §6.2.

	Correct	Partial	Incorrect	Ungrammatical	Is Logical?
Direct Prompting	28.7±3.7	20.0±2.5	38.8±8.7	12.5±2.5	62.0±0.5
MURMUR	**55.0±2.5**	1.2±1.2	38.8±3.7	5.0±2.5	**95.4±0.2**

Table 7: Human evaluation of logical correctness for LogicNLG. 'Is Logical' denotes the percentage of correct generations that also involve some underlying logical computations.

Results. Table 6 shows the results on the test set of LogicNLG. MURMUR significantly improves upon DP and CoT prompting by up to 2.4 points in BLEU-3 ($p < 0.001$). We attribute this to two factors: (1) leveraging symbolic modules for logical skills that ensure their correctness, (2) delegating the task of converting a path to natural language to an LLM. Both CoT and BART+SR, while generating intermediate reasoning paths, do not use executable modules and hence cannot guarantee valid compositionality or logical correctness of the reasoning steps. MURMUR also improves upon the supervised Field-Infusing model. Finally, MURMUR obtains some improvement with the saliency metric, indicating that it helps in choosing more salient paths. Refer to Appendix C for studies on the number and variation in demonstrations.

6.2 Human Evaluation of Logical Correctness

Next, we conduct human evaluation to compare the logical correctness of the generations from DP and MURMUR. Two NLP experts annotate 40 randomly chosen generations from eight different tables. In particular, they take part in two studies. First, they classify each generation into whether it is (a) ungrammatical, (b) grammatical but incorrect, (c) grammatical but partially correct, or (d) grammatical and also fully correct. Next, for each fully correct generation, they annotate whether it involves any underlying logical operation (like counting, summation, etc) or are mere surface realizations of the table content. We observe from Table 7 that MURMUR not only generates 26% more correct outputs, but about 95% of those generations also involve some logical operations. In summary, MURMUR is most beneficial in two scenarios: (1) generations that require many steps of reasoning, (2) generations that require logical reasoning. The first capability comes from the fact that MURMUR is specifically designed to compose multiple steps of reasoning through its grammar and value functions. The second benefit is because of the presence of symbolic modules that ensure logical correctness. These two capabilities are specifically required in long complex tables involving numerical columns where there is a need to summarize content (e.g., by filtering, averaging a numerical column, etc). Generating reasoning paths through logical modules ensures that almost all generations are logical derivations from the table, an ability that is significantly harder to achieve through direct prompting. See Fig. 8 in the appendix for an illustrative example of the generations of MURMUR for long complex tables.

7 Related Work

Multi-step Reasoning over Text. Recent developments in LLMs (Brown et al., 2020; Zhang et al., 2022; Thoppilan et al., 2022; Chowdhery et al., 2022) have enabled significant progress in few-shot methods for logical reasoning tasks (Wei et al., 2022; Creswell et al., 2022; Nye et al., 2021; Wang et al., 2022c; Zelikman et al., 2022; Zhou et al., 2022; Dasgupta et al., 2022; Kojima et al., 2022; Dohan et al., 2022). Representative methods like CoT prompting output intermediate reasoning steps before generating the final answer. However, the reasoning steps are all generated in one go from a single model, potentially leading to unfaithful reasoning due to the lack of explicit conditioning between the steps (Creswell and Shanahan, 2022). MURMUR overcomes this issue by developing granular modules that are capable of performing specialized skills by *explicitly* conditioning on the outputs from previous reasoning steps. Conceptually, MURMUR bears similarity with the Selection-Inference modular architecture (Creswell et al., 2022; Creswell and Shanahan, 2022). However, their focus is on QA and reasoning over textual context (Saha et al., 2020, 2021b; Tafjord et al., 2021; Dalvi et al., 2021; Bostrom et al., 2022). A few concurrent works have also proposed neuro-symbolic approaches for reasoning over text (Gao et al., 2022; Wang et al., 2022b; Chen et al., 2022; Cheng et al., 2022). Different from these, we tackle a more challenging setup of multi-step reasoning for *controlled generation* from *semi-structured data*.

Modular Reasoning over Text. Neural Module Networks learn and execute compositional programs over modules (Andreas et al., 2016; Jiang and Bansal, 2019; Gupta et al., 2020; Subramanian

et al., 2020; Saha et al., 2021a). While their modules typically output attention maps, prior works have also used text-in text-out modules whose input/output data types are *strings* (Khot et al., 2021, 2022; Saha et al., 2022). MURMUR's modules are a generalization of text-in text-out modules since they can capture operations involving complex data types (like *tables*) and *strings*, among others. The data to text transition is also clearly represented through the compositions of our modules, unlike attention maps-based modules whose interpretability has often been debated (Serrano and Smith, 2019).

Data-to-Text Generation. Existing methods for data-to-text generation include (1) supervised methods that finetune seq2seq LMs (Kale and Rastogi, 2020; Chen et al., 2020b; Ribeiro et al., 2021; Ke et al., 2021; Xiang et al., 2022), (2) pipeline modular methods (Reiter and Dale, 1997; Reiter, 2007; Laha et al., 2020; Kasner and Dušek, 2022), and (3) few-shot methods that assume access to a large corpus of unlabeled examples for data augmentation or retrieving similar examples (Puduppully et al., 2019; Zhao et al., 2020; Trisedya et al., 2020; Su et al., 2021). Unlike prior modular methods, MURMUR uses few-shot neural or symbolic modules. Unlike past few-shot methods, MURMUR works well with as few as one demonstration, without requiring access to any unlabeled corpus.

8 Conclusion

We presented MURMUR, a neuro-symbolic modular reasoning approach for data-to-text generation. MURMUR shows the benefits of building interpretable modular text generation systems by breaking a task down into sub-problems and then solving them through separate modules, without requiring module-specific supervision. It utilizes the power of LLMs in solving linguistic sub-tasks through in-context learning while delegating the logical sub-tasks to symbolic modules. MURMUR generalizes the concept of modules by treating them as functions and defining their behaviors through expected input/output data types and compositions with a grammar (analogous to function compositions).

Limitations

MURMUR relies on large language models for few-shot linguistic skills like surface realization and text fusion. It is probable that smaller models do not work as well, in which case one may curate additional training data to train these modules. We also note that our choice of logical modules is motivated by the characteristics of the task. Hence, it is conceivable that other data-to-text generation tasks might benefit from incorporating additional modules. MURMUR does not make any assumptions about the type or implementation of the modules and it should be straightforward to extend our method to other data-to-text generation tasks.

We limit our experiments to English datasets. We also adopt a simple prompting strategy for converting a reasoning path to a natural language summary by representing the path as a string. This works well in practice and OPT is typically able to resolve the module names and their arguments correctly. However, more future work is needed to understand when this fails so that better prompting methods can be developed. Despite the known limitations of standard automatic metrics like BLEU and METEOR, we use them to compare our method to previous works. While this is not ideal, we have performed comprehensive human evaluation for both tasks to further verify our claims.

Ethics Statement

Large Language Models can be prone to generate toxic and unwanted content (Weidinger et al., 2021). Since MURMUR uses focused modules to accomplish specific skills, we believe that this might help limit inadvertent negative impacts. Furthermore, the presence of specific modules should provide users with more trust and control in real-world scenarios, allowing one to verify, debug, and improve the capabilities of these modules.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

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

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction through search over statement compositions. *arXiv preprint arXiv:2201.06028*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020b. Kgpt: Knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8635–8648.

Wenqing Chen, Jidong Tian, Yitian Li, Hao He, and Yaohui Jin. 2021. De-confounded variational encoder-decoder for logical table-to-text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5532–5542.

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020c. Logic2text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2096–2111.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2022. Binding language models in symbolic languages. *arXiv preprint arXiv:2210.02875*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A Saurous, Jascha Sohl-dickstein, et al. 2022. Language model cascades. *arXiv preprint arXiv:2207.10342*.

Ondřej Dušek and Filip Jurcicek. 2015. Training a natural language generator from unaligned data. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461.

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.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

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.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020. Neural module networks for reasoning over text. In *ICLR*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Yichen Jiang and Mohit Bansal. 2019. Self-assembling modular networks for interpretable multi-hop reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4474–4484.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102.

Zdeněk Kasner and Ondřej Dušek. 2022. Neural pipeline for zero-shot data-to-text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3914–3932.

Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2526–2538.

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. Text modular networks: Learning to decompose tasks in the language of existing models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1264–1279.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Anirban Laha, Parag Jain, Abhijit Mishra, and Karthik Sankaranarayanan. 2020. Scalable micro-planned generation of discourse from structured data. *Computational Linguistics*, 45(4):737–763.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213.

Kathleen McKeown. 1992. *Text generation*. Cambridge University Press.

Hongyuan Mei, Mohit Bansal, and Matthew R Walter. 2016. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. In *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 201–206. Association for Computational Linguistics.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

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.

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

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, 01, pages 6908–6915.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *proceedings of the eleventh European workshop on natural language generation (ENLG 07)*, pages 97–104.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227.

Amrita Saha, Shafiq Joty, and Steven CH Hoi. 2021a. Weakly supervised neuro-symbolic module networks for numerical reasoning. *arXiv preprint arXiv:2101.11802*.

Swarnadeep Saha, Sayan Ghosh, Shashank Srivastava, and Mohit Bansal. 2020. PRover: Proof generation for interpretable reasoning over rules. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 122–136.

Swarnadeep Saha, Prateek Yadav, and Mohit Bansal. 2021b. multiPRover: Generating multiple proofs for improved interpretability in rule reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3662–3677.

Swarnadeep Saha, Shiyue Zhang, Peter Hase, and Mohit Bansal. 2022. Summarization programs: Interpretable abstractive summarization with neural modular trees. *arXiv preprint arXiv:2209.10492*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909.

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. 2020. Obtaining faithful interpretations from compositional neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5594–5608.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Bayu Trisedya, Jianzhong Qi, and Rui Zhang. 2020. Sentence generation for entity description with content-plan attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 9057–9064.

Fei Wang, Zhewei Xu, Pedro Szekely, and Muhao Chen. 2022a. Robust (controlled) table-to-text generation with structure-aware equivariance learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5037–5048, Seattle, United States. Association for Computational Linguistics.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022b. Behavior cloned transformers are neurosymbolic reasoners. *arXiv preprint arXiv:2210.07382*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022c. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Jiannan Xiang, Zhengzhong Liu, Yucheng Zhou, Eric P. Xing, and Zhiting Hu. 2022. ASDOT: Any-shot data-to-text generation with pretrained language models. In *EMNLP Findings*.

Kaiyu Yang, Jia Deng, and Danqi Chen. 2022. Generating natural language proofs with verifier-guided search. In *EMNLP*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Module Name	Input Data Type	Output Data Type	Description
filter_eq filter_not_eq	table, string, string	table	Returns a table with the rows where entry in the input column (second argument) is equal or not equal to the input value (third argument).
filter_greater filter_greater_eq filter_lesser filter_lesser_eq	table, string, number	table	Returns a table with the rows where a numerical column (second argument) is greater than or less than (or equal to) the input number (third argument).
filter_all	table, string	table	Returns the whole table.
arg_max arg_min	table, string	row	Returns the row with the minimum or maximum value for the input column (second argument).
max min avg sum	table, string	number	Returns the maximum, minimum, average or sum of numbers in the input column (second argument).
count	table	number	Returns the number of rows in the table.
all_eq all_not_eq	table, string, string	bool	Returns whether all entries in the input column are equal (or not equal to) the input value.
all_greater all_less all_greater_eq all_less_eq	table, string, number	bool	Returns whether all entries in the input column are greater than or less than (or equal to) the input number.
most_eq most_not_eq	table, string, string	bool	Returns whether most entries in the input column are equal (or not equal to) the input value.
most_greater most_less most_greater_eq most_less_eq	table, string, number	bool	Returns whether most entries in the input column are greater than or less than (or equal to) the corresponding number.
only	table	bool	Returns whether the table has exactly one row.
hop	row, string	string	Returns the entry corresponding to the input column in the row.
eq	string, string	bool	Returns whether the two inputs are equal or not.

Table 8: List of modules for LogicNLG with their corresponding input / output data types and descriptions.

A Modules for Table-to-Text Generation (Cont. from §3.1)

Table 8 shows the list of all modules for LogicNLG. Our choice of modules is motivated from prior work (Chen et al., 2020c) that defines similar modules for generating logical summaries from open-domain tables.

B Additional Experiments on WebNLG (Cont. from §5)

B.1 Effect of Number of Demonstrations

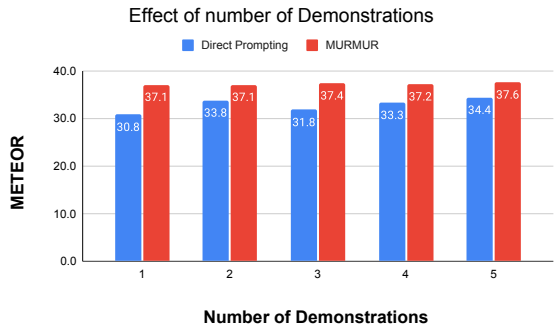

Figure 3: Comparison of METEOR scores for Direct Prompting (DP) versus MURMUR with varying number of demonstrations on WebNLG test set. DP shows improved performance with more demonstrations while MURMUR's gains are marginal.

In Fig. 3, we compare the METEOR scores of DP and MURMUR by varying the number of demonstrations. DP shows improved performance with more demonstrations, while MURMUR's improvements are marginal. In the process of providing more demonstrative examples, DP implicitly learns the underlying step-wise reasoning process, while such phenomenon is explicitly captured through one demonstration in MURMUR.

B.2 Effect of Variations of Demonstrations

	BLEU	METEOR
Direct Prompting (k=1)	31.1±0.5	29.8±0.1
Direct Prompting (k=5)	38.3±0.4	33.6±0.1
MURMUR (k=1)	**40.1±0.3**	**37.1±0.5**

Table 9: Comparison of different few-shot methods on the WebNLG validation set. We report mean and variance of BLEU and METEOR scores with three different random seeds for choosing demonstrations from the training set.

In Table 9, we compare the performance of few-shot baselines on the validation set of WebNLG and analyze the effect of different choices of random demonstrations on in-content learning. Using three different random seeds, we show that all methods are fairly robust to randomness in demonstrations.

C Additional Experiments on LogicNLG (Cont. from §6)

C.1 Effect of Number of Demonstrations

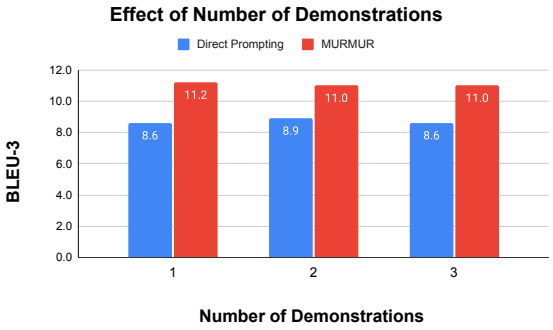

Figure 4: Comparison of BLEU-3 scores for Direct Prompting versus MURMUR with varying number of demonstrations on LogicNLG validation set. For both methods, results do not improve further with more demonstrations.

In Fig. 4, we compare BLEU-3 scores of DP and MURMUR by varying the number of demonstrations from 1 to 3. Unlike WebNLG, we do not observe any noticeable improvements in in-context learning capabilities with more demonstrations, possibly because of the inherent difficulty of generating logical summaries from tables.

C.2 Effect of Variations in Demonstrations

In Table 10, we study the effect of randomness in the choice of a single demonstration for LogicNLG. We report mean and variance of BLEU scores for each method with a randomly chosen demonstration from the training examples. Similar to WebNLG, all methods are fairly robust to the choice of demonstrations and exhibit comparable variance in performance.

C.3 Effect of Different Beam Sizes in Best-first Search of MURMUR

At each step of the search, MURMUR keeps track of the highest scoring reasoning paths. Table 11 compares the effect of the beam size for our search algorithm on the LogicNLG validation set. Perhaps unsurprisingly, maintaining a bigger beam i.e., conducting a more exhaustive search leads to some

	BLEU-1 / BLEU-2 / BLEU-3
Direct Prompting (k=1)	37.0±0.2 / 18.9±0.1 / 8.5±0.1
COT Prompting (k=1)	36.5±0.1 / 18.9±0.1 / 8.7±0.3
MURMUR (k=1)	**40.5±0.1 / 22.2±0.0 / 10.8±0.1**

Table 10: Comparison of different few-shot methods on the LogicNLG validation set. We report mean and variance of BLEU scores with two random seeds for choosing one demonstration from the training set.

Beam Size	BLEU-1 / BLEU-2 / BLEU-3
10	39.7 / 21.2 / 10.3
20	40.2 / 21.8 / 10.7
50	40.5 / 22.2 / 10.8
100	**40.7 / 22.5 / 10.9**

Table 11: Effect of beam size in MURMUR's search algorithm on BLEU scores of LogicNLG validation set.

improvements in BLEU scores, however, the gain mostly saturates with beam sizes of around 50-100.

C.4 Further Analysis of Saliency Metric (Cont. from §4.2)

Training Data Construction. In Fig. 5, we show an illustrative example of the training data creation process for our saliency metric. In 'Incorrect Partial Path-1', when we perturb the *avg* module with the *sum* module, we aim to teach the model that although both are valid reasoning steps, averaging over the column 'points' is a more salient and informative reasoning step than summing over the column 'year'. Similarly, in 'Incorrect Partial Path-2', when we perturb the input to the module *avg* by performing average over the column 'wins', we want the model to learn the salient columns to reason over for a given module.

Effect of Varying Supervision on Metric Accuracy and Downstream Performance. We conduct an in-depth analysis of the saliency metric used to score the reasoning steps in MURMUR. As shown in Table 12, we vary the amount of supervision for training the saliency metric and study its effect on the validation set accuracy (in identifying whether a partial reasoning path is correct or not) and also on the downstream LogicNLG BLEU scores. Our key takeaway is that a small number of gold reasoning paths (about 200, spanning 100 tables) is enough to train a good saliency metric that not only achieves a high classification accuracy of 76% but also leads to a BLEU-3 score of 10.8 on LogicNLG. Increasing the training data further to 7k gold paths (equivalently, 42k correct and incorrect partial paths) increases the classification accuracy to 82% but does not impact LogicNLG performance much.

# Gold Tables	# Gold Paths	# Samples (Pos/Neg/ All)	Acc.	LogicNLG BLEU-3
100	221	769/729/1498	76.16	10.8
200	443	1534/1457/2991	78.52	10.6
500	1085	3773/3633/7406	80.32	10.9
3000	7145	21.5k/20.5k/42.0k	82.84	10.7

Table 12: Effect of varying amount of supervision for the saliency metric on the metric accuracy (Acc.) and on downstream LogicNLG BLEU scores. Metric accuracy is computed on 4.4k validation samples consisting of 2264 correct paths (positive samples) and 2179 incorrect paths (negative samples).

D Prompts (Cont. from §4)

WebNLG. Table 13 shows an example of direct prompting (Zhang et al., 2022) for WebNLG. In Table 15 and Table 16, we show the prompts for the *surface realization* and *text fusion* modules in MURMUR. Note that the single demonstration for direct prompting is decomposed into individual reasoning steps for the two modules in MURMUR.

```
Let's convert triples to sentences
###
Triples: A.S._Gubbio_1910 | league | Serie_D # Italy | leader | Pietro_Grasso # Italy | capital | Rome # A.S._Gubbio_1910 | ground | Italy # Serie_D | champions | S.S._Robur_Siena
Output: S.S. Robur Siena are champions of Serie D in which AS Gubbio 1910 also play. This latter club have their home ground in Italy where the capital city is Rome and the leader is Pietro Grasso.
###
Triples: {triples}
Output:
```

Table 13: Example of Direct Prompting for WebNLG.

```
Let's convert triples to sentences step-by-step
###
Triples: A.S._Gubbio_1910 | league | Serie_D # Italy | leader | Pietro_Grasso # Italy | capital | Rome # A.S._Gubbio_1910 | ground | Italy # Serie_D | champions | S.S._Robur_Siena
Output: AS Gubbio 1910 plays in Serie D. # Pietro Grasso is the leader of Italy. # ... # S.S. Robur Siena are champions of Serie D in which AS Gubbio 1910 also play. This latter club have their home ground in Italy where the capital city is Rome and the leader is Pietro Grasso.
###
Triples: {triples}
Output:
```

Table 14: Example of Chain-of-Thought Prompting for WebNLG. The intermediate reasoning steps (truncated for clarity) are concatenated together and we consider the last step as the final summary.

Table Topic: Reinhold Roth

Year	Class	Team	Points	Wins
1979	350cc	yamaha	3	0
1980	250cc	yamaha	4	0
1982	250cc	yamaha	4	0
1982	500cc	suzuki	0	0
1983	250cc	yamaha	14	0
1984	500cc	honda	14	0
1985	250cc	romer-juchem	29	0
1986	250cc	hb - honda	10	0
1987	250cc	hb - honda	108	1
1988	250cc	hb - honda	158	0
1989	250cc	hb - honda	190	2
1990	250cc	hb - honda	52	0

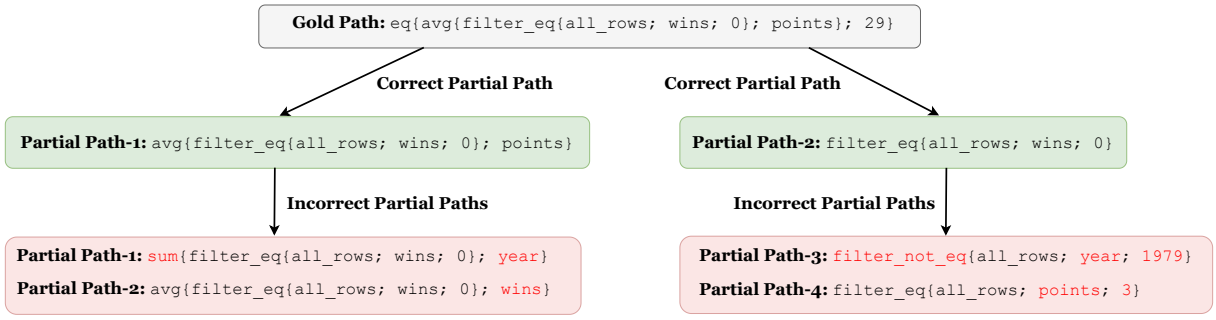

Figure 5: Training data creation process for the saliency metric. Given a gold path, we create correct (positive) partial paths by breaking the path at each step. From each correct partial path, we create incorrect partial paths by doing two kinds of perturbations, one at the module level and another at the inputs level (Cont. from §4.2).

Let's convert a triple to a sentence
###
Triple: A.S._Gubbio_1910 | league | Serie_D
Sentence: AS Gubbio 1910 plays in Serie D.
###
Triple: Italy | leader | Pietro_Grasso
Sentence: Pietro Grasso is the leader of Italy.
###
Triple: Italy | capital | Rome
Sentence: Rome is the capital of Italy.
###
Triple: A.S._Gubbio_1910 | ground | Italy
Sentence: Italy is the home ground of AS Gubbio 1910.
###
Triple: Serie_D | champions | S.S._Robur_Siena
Sentence: S.S. Robur Siena are champions of Serie D.
###
Triple: {triple}
Sentence:

Table 15: Example of the Surface Realization prompt for MURMUR in WebNLG.

LogicNLG. Table 17 shows an example of direct prompting for LogicNLG. Table 18 shows an example prompt for the *surface realization* module in LogicNLG. We only provide the table topic, table header, and the reasoning path in the prompt. We do not add the table content to the prompt because all the information needed by the model to generate the summary is typically present in the reasoning path. Any other contextual information about the table can also be inferred from the table header and topic. We observe that adding the table content makes the model more prone to hallucinations because it may not limit its generation to the information provided in the reasoning path alone.

E Examples of MURMUR Summaries

In Fig. 6, we show representative examples of summaries generated by Direct Prompting and MURMUR for WebNLG. Fig. 7 shows the step-wise summary generation process of MURMUR for WebNLG. In Fig. 8 and 9, we show representative examples of the reasoning paths and summaries generated for two tables in LogicNLG.

Input	Arlington,_Texas | isPartOf | Texas # Texas | largestCity | Houston # Texas | language | English_language
Direct Prompting	Texas is the second largest state in the United States of America. Its capital is Austin and the largest city is Houston. (**Omissions = 2**)
MURMUR	Houston is the largest city in Texas where English is the official language and Arlington , Texas is a city in Texas. (**Omissions = 0**)
Input	Hays_County_Texas | countySeat | San_Marcos,_Texas # Austin,_Texas | isPartOf | Hays_County,_Texas # Texas | language | Spanish_language # Austin,_Texas | isPartOf | Texas # Texas | largestCity | Houston
Direct Prompting	Hays County, Texas is a county in the U.S. state of Texas. (**Omissions = 4**)
MURMUR	Spanish is the official language of Texas. San Marcos is the county seat of Hays County, Texas where Houston is the largest city in Texas where Austin, Texas is part of Texas. (**Omissions = 0**)
Input	Vermont | largestCity | Burlington_Vermont # Alvah_Sabin | region | Vermont # Alvah_Sabin | activeYearsEndDate | 1857-03-03
Direct Prompting	Alvah Sabin was born on March 3, 1857 in Vermont. (**Omissions = 2**)
MURMUR	Burlington is the largest city in Vermont where Alvah Sabin is from and he played from 1857-03-03 to 1857-03-03. (**Omissions = 0**)

Figure 6: Examples of summaries generated by Direct Prompting and MURMUR for WebNLG. Hallucinations are marked in red, omissions are marked in olive, and disfluencies are marked in blue. Omission count of triples is shown in brackets next to the generations.

```
Let's combine two sentences
###
First Sentence: S.S. Robur Siena are champions of Serie D.
Second Sentence: AS Gubbio 1910 plays in Serie D.
Combined Sentence: S.S. Robur Siena are champions of Serie D in which AS Gubbio 1910 also play.
###
First Sentence: Rome is the capital of Italy.
Second Sentence: Pietro Grasso is the leader of Italy.
Combined Sentence: Rome is the capital of Italy where Pietro Grasso is the leader.
###
First Sentence: S.S. Robur Siena are champions of Serie D in which AS Gubbio 1910 also play.
Second Sentence: Italy is the home ground of AS Gubbio 1910.
Combined Sentence: S.S. Robur Siena are champions of Serie D in which AS Gubbio 1910 also play. This latter club have their home ground in Italy.
###
First Sentence: S.S. Robur Siena are champions of Serie D in which AS Gubbio 1910 also play. This latter club have their home ground in Italy.
Second Sentence: Rome is the capital of Italy where Pietro Grasso is the leader.
Combined Sentence: S.S. Robur Siena are champions of Serie D in which AS Gubbio 1910 also play. This latter club have their home ground in Italy where the capital city is Rome and the leader is Pietro Grasso.
###
First Sentence: {sent1}
Second Sentence: {sent2}
Combined Sentence:
```

Table 16: Example of the Text Fusion prompt for MURMUR in WebNLG.

```
Let's generate a logically entailed statement from the table
###
Table Topic: 1938 U.S. Open (golf)
Table Header: place # player # country # score # to par # money
Table Content: 1 # ralph guldahl # united states # 74 + 70 + 71 + 69 = 284 # e # 1000 | ... | 10 # gene sarazen # united states # 74 + 74 + 75 + 73 = 296 # + 12 # 106
Generation: The majority of the players in the 1938 US Open scored at least 9 over par or above .
###
Table Topic: {table_topic}
Table Header: {table_header}
Table Content: {table_content}
Generation:
```

Table 17: Example of Direct Prompting for LogicNLG. Each row in the table is separated by a 'l' and each entry in a row is separated by a '#'. The table content is truncated for conciseness.

```
Let's generate a logically entailed statement from the table for the reasoning path
###
Table Topic: 1938 U.S. Open (golf)
Table Header: place # player # country # score # to par # money
Reasoning Path: most_greater_eq { all_rows ; to par ; 9 }
Generation: The majority of the players in the 1938 US Open scored at least 9 over par or above .
###
Table Topic: {table_topic}
Table Header: {table_header}
Reasoning Path: {reasoning_path}
Generation:
```

Table 18: Example of Surface Realization Prompt for MURMUR in LogicNLG.

Input Graph: Antwerp_International_Airport | operatingOrganisation | Flemish_Government # Antwerp_International_Airport | elevationAboveTheSeaLevel_(in_metres) | 12.0 # Antwerp_International_Airport | owner | Flemish_Region # Antwerp_International_Airport | runwayLength | 600.0

#Step	Module	Input	Output
1	Surface Realization	Antwerp_International_Airport | runwayLength | 600.0	The runway at Antwerp International Airport is 600.0 metres long.
2	Surface Realization	Antwerp_International_Airport | elevationAboveTheSeaLevel_(in_metres) | 12.0	Antwerp International Airport is 12.0 metres above sea level.
3	Surface Realization	Antwerp_International_Airport | operatingOrganisation | Flemish_Government	Antwerp International Airport is operated by the Flemish Government.
4	Text Fusion	Antwerp International Airport is owned by the Flemish Region.; Antwerp International Airport is operated by the Flemish Government.	Antwerp International Airport is owned by the Flemish Region and operated by the Flemish Government.
5	Text Fusion	Antwerp International Airport is 12.0 metres above sea level.; Antwerp International Airport is owned by the Flemish Region and operated by the Flemish Government.	Antwerp International Airport is 12.0 metres above sea level and is owned by the Flemish Region and operated by the Flemish Government.
6	Text Fusion	The runway at Antwerp International Airport is 600.0 metres long.; Antwerp International Airport is 12.0 metres above sea level and is owned by the Flemish Region and operated by the Flemish Government.	The runway at Antwerp International Airport is 600.0 metres long and is 12.0 metres above sea level and is owned by the Flemish Region and operated by the Flemish Government.

Figure 7: Illustration of the step-wise summary generation process of MURMUR for WebNLG. Each step consists of a module (Surface Realization or Fusion), the input to the module (a triple or a pair of texts) and the output summary (Cont. from §3.3).

Table Topic: 2002 Colorado Rockies Season

Date	Opponent	Score	Loss	Attendance	Record
June 1	giants	5 - 4	hernández (5 - 5)	40893	30 - 26
June 2	giants	9 - 2	thomson (6 - 4)	40651	30 - 27
June 3	dodgers	11 - 5	jiménez (1 - 3)	30150	30 - 28
June 4	dodgers	10 - 4	jones (0 - 1)	30195	30 - 29
June 5	dodgers	8 - 6	daal (4 - 2)	31793	31 - 29
June 7	blue jays	8 - 0	hampton (3 - 7)	20032	31 - 30
June 8	blue jays	3 - 1	thomson (6 - 5)	21298	31 - 31
June 9	blue jays	3 - 2	jiménez (1 - 4)	20328	31 - 32
June 10	red sox	7 - 3	neagle (4 - 3)	33508	31 - 33
June 11	red sox	3 - 1	fossum (2 - 1)	32340	32 - 33
June 12	red sox	7 - 5	hampton (3 - 8)	31583	32 - 34
June 14	indians	5 - 3	thomson (6 - 6)	40156	32 - 35
June 15	indians	7 - 4	paronto (0 - 2)	41870	33 - 35
June 16	indians	5 - 4	neagle (4 - 4)	40792	33 - 36
June 18	yankees	10 - 5	jennings (8 - 3)	48738	33 - 37
June 19	yankees	20 - 10	white (1 - 5)	48821	33 - 38
June 20	yankees	14 - 11 (10)	karsay (3 - 3)	48916	34 - 38
June 21	devil rays	8 - 7 (10)	yan (3 - 3)	30284	35 - 38
June 22	devil rays	6 - 5 (11)	kent (0 - 2)	31190	36 - 38
June 23	devil rays	6 - 5	kennedy (5 - 6)	31043	37 - 38
June 24	dodgers	4 - 1	ishii (11 - 3)	34641	38 - 38
June 25	dodgers	4 - 0	thomson (6 - 7)	23635	38 - 39
June 26	dodgers	5 - 3	chacón (3 - 5)	25083	38 - 40
June 27	dodgers	7 - 1	neagle (4 - 5)	41279	38 - 41
June 28	mariners	6 - 2	jennings (8 - 4)	45118	38 - 42
June 29	mariners	8 - 1	hampton (4 - 9)	45790	38 - 43
June 30	mariners	4 - 3	sasaki (2 - 2)	45928	39 - 43

Correct Summaries

Reasoning Path-1: eq { hop { argmax { all_rows ; attendance } ; date } ; june 20 }

Summary-1: The game on June 20 drew the highest attendance of the 2002 Colorado Rockies season .

Reasoning Path-2: eq { count { filter_eq { all_rows ; opponent ; indians } } ; 3 }

Summary-2: The Colorado Rockies played against the indians 3 times in the 2002 season .

Reasoning Path-3: eq { hop { argmin { all_rows ; attendance } ; loss } ; hampton (3 - 7) }

Summary-3: Hampton (3 - 7) was the game with the lowest attendance in the 2002 Colorado Rockies season .

Reasoning Path-4: eq { count { all_rows } ; 27 }

Summary-4: The Colorado Rockies played 27 games in the 2002 season .

Reasoning Path-5: eq { max { all_rows ; attendance } ; 48916 }

Summary-5: The highest attendance for the 2002 colorado rockies season was 48916 .

Figure 8: Sample table from LogicNLG dataset and five diverse logical summaries generated by MURMUR. Each color code in the table cells highlights parts of the table relevant to a MURMUR summary. All generated summaries are logically correct.

Table Topic: List of Latvian submissions for the academy award for best foreign language film

year (ceremony)	film title used in nomination	original title	director	result
1992 (65th)	the child of man	cilvēka bērns	jānis streičs	not nominated
2008 (81st)	defenders of riga	rīgas sargi	aigars grauba	not nominated
2010 (83rd)	hong kong confidential	amaya	māris martinsons	not nominated
2012 (85th)	gulf stream under the iceberg	golfa straume zem ledus kalna	yevgeni pashkevich	not nominated
2013 (86th)	mother , i love you	mammu , es tevi mīlu	jānis nords	tbd

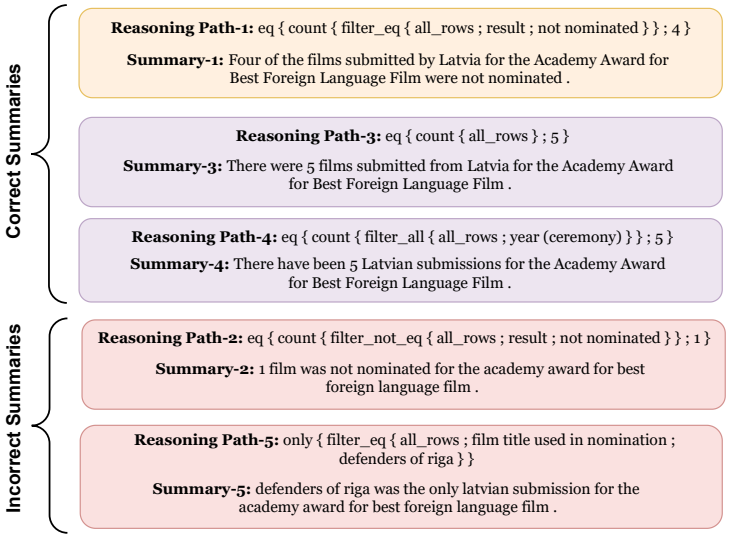

Figure 9: Sample table from LogicNLG dataset and five diverse logical summaries generated by MURMUR. Each color code in the table cells highlights parts of the table relevant to a MURMUR summary. The red marked blocks are incorrect summaries generated by MURMUR.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
9

☑ A2. Did you discuss any potential risks of your work?
10

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

5, 6

☑ B1. Did you cite the creators of artifacts you used?
5, 6

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
5, 6

C ☑ Did you run computational experiments?

5, 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4, 5, 6

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5, 6, Appendix B, C

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5, 6

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
5, 6

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
5, 6

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.