The Whole Truth and Nothing But the Truth: 
Faithful and Controllable Dialogue Response Generation 
with Dataflow Transduction and Constrained Decoding

Hao Fang∗ Anusha Balakrishnan∗ Harsh Jhamtani∗
John Bufe Jean Crawford Jayant Krishnamurthy
Adam Pauls Jason Eisner Jacob Andreas Dan Klein
Microsoft Semantic Machines <sminfo@microsoft.com>

Abstract

In a real-world dialogue system, generated text must be truthful and informative while remaining fluent and adhering to a prescribed style. Satisfying these constraints simultaneously is difficult for the two predominant paradigms in language generation: neural language modeling and rule-based generation. We describe a hybrid architecture for dialogue response generation that combines the strengths of both paradigms. The first component of this architecture is a rule-based content selection model defined using a new formal framework called dataflow transduction, which uses declarative rules to transduce a dialogue agent’s actions and their results (represented as dataflow graphs) into context-free grammars representing the space of contextually acceptable responses. The second component is a constrained decoding procedure that uses these grammars to constrain the output of a neural language model, which selects fluent utterances. Our experiments show that this system outperforms both rule-based and learned approaches in human evaluations of fluency, relevance, and truthfulness.

1 Introduction

In a task-oriented dialogue system, response generation is naturally posed as a conditional language modeling problem: dialogue agents must produce a contextually appropriate natural language string conditioned on the history of the user and agent interaction. But unlike many language generation problems, a good dialogue response generation model is not (just) a model of typical human utterances in context. Instead, effective dialogue agents must balance fluent generation with a set of much stricter constraints.

Consider the dialogue shown in Fig. 1. In turn (1) of this dialogue, the user makes a request, which the dialogue agent correctly translates into a computation—here represented as a dataflow graph (Fig. 1a) in the style of Semantic Machines et al. (2020). The agent now needs to accurately describe this computation’s return value (namely, 5). The wrong answer in Fig. 1b shows it instead describing a different value that happens to appear elsewhere in the dataflow graph. Turn (2) illustrates a more subtle risk: due to a speech recognition error, the agent has mistakenly created a meeting with Tara Smith rather than Sarah Smith. The wrong answer in Fig. 1c shows it describing this result too briefly, which might lead the user to assume that their request was completed successfully. To avoid confusion, a system designer might wish to ensure that the agent instead echoes back
to the user the details of the agent’s action.

This example highlights challenges in building real-world dialogue response generation systems.

First, response generation is not simply a problem of describing the result of a computation in natural language. In some cases, response generators may also usefully describe the provenance of that result—the computation itself and its intermediate values. In many human-to-human conversations, a response as detailed as Fig. 1c would be over-informative, violating Grice’s maxim of quantity (1975). But for a speaker that is prone to mistakes, such as an AI agent, describing its own understanding can increase user trust when the understanding is accurate and provides an opportunity for correction when it is not.

Second, dialogue response generation systems must guarantee truthfulness: since the user often has no way to check the responses, even occasional errors could have disastrous consequences and would greatly undermine trust. Yet truthful utterances might be low-probability under a domain-general language model (LM), particularly when they reflect errors in language understanding (as in Fig. 1c).

Finally, response generation systems must support declarative specification of agent behavior. When confusing or felicitous responses are discovered, it should be possible to easily and precisely modify them without changing the dialogue agent’s behavior in other contexts.

In recent years, the main focus of academic dialogue research has been on “end-to-end” learned models for response generation, especially neural sequence models (Vinyals and Le, 2015; Zhang et al., 2020b). But while such models excel at producing fluent and coherent output, research continues to find that they struggle in maintaining faithfulness (Wiseman et al., 2017; Maynez et al., 2020; Raunak et al., 2021; Liu et al., 2023; Zhang et al., 2023). Perhaps more fundamentally, because the behavior of such systems is encoded implicitly in their training data, designing a dialogue system requires system builders to write and edit a large number of training examples whose final effect may be difficult to predict.

As a result, many dialogue systems in the real world remain rule-based: system builders hand-write rules (e.g., in the form of a synchronous grammar) for transforming dialogue states into text, and these rules are applied directly during deployment. But such rule-based systems are also notoriously difficult to build and maintain (Walker et al., 2002; Reiter, 2022). They require designers to anticipate every low-level question about surface realization, and to encode these in the same grammar that is responsible for enforcing high-level properties like truthfulness.

Given the many strengths of modern LMs, is there a way to leverage them while satisfying the numerous other demands on dialogue response generation systems? In this paper, we describe a hybrid approach that combines the advantages of end-to-end and rule-based approaches. This approach has two components:

- A dataflow transduction procedure (§3) that maps any computation by the agent (represented as a dataflow graph) to a small context-free grammar (CFG) that defines the space of natural language descriptions or responses allowed for the given computation. The mapping is defined by declarative rules. This formal framework makes it possible to write rules to precisely and truthfully describe both data and its provenance, while performing supplementary computation where needed to produce informative responses.

- A constrained decoding procedure (§4) that uses beam search to identify strings that are both grammatical under the CFG and probable under a given language model (LM). In effect, this intersects the CFG with the LM.

This makes it possible to decompose language generation into a content selection model (implemented by the dataflow transducer) and a separate fluency model (implemented by the LM). Hybrid generation systems of this kind have a long history in NLP, dating back to Knight and Hatzivas-siloglou (1995) and Langkilde and Knight (1998). They mapped an abstract meaning representation (AMR) to an acyclic finite-state automaton (FSA) and scored its paths with an n-gram LM. We replace AMR with dataflow, replace their mapping rules with dataflow transduction rules, upgrade their FSA to a CFG, and upgrade their n-gram LM to a neural LM. In this way, we respectively support computation graphs, arbitrary tests and transductions, nested syntactically typed generation templates (already present in Knight and Hatzivas-siloglou, 1995), and modern language models.
Together, dataflow transduction and constrained decoding allow a compact generation system to faithfully and fluently describe a complex and open-ended space of actions. We built such a hybrid system for calendar event queries in the SMCalFlow domain (Semantic Machines et al., 2020). When evaluated on a subset of annotated dialogues, it was consistently rated as more truthful, relevant, and fluent than either a rule-based or end-to-end neural system (§5.2). Results were similar on MultiWOZ dialogues (Budzianowski et al., 2018; Eric et al., 2020) (§5.4). Code, data, and trained models used in our experiments are released at https://github.com/microsoft/dataflow2text.

2 Problem Formulation

We study the problem of response generation for task-oriented dialogue. A dialogue, like the one in Fig. 1, consists of a sequence of turns $k$, each consisting of a user utterance $x_k$, one or more actions $a_k$, and an agent response $y_k$. The job of a dialogue agent is to predict an appropriate action and response from a dialogue history, i.e.,, to map from $(x_1, a_1, y_1, x_2, a_2, y_2, \ldots, x_n) \mapsto (a_n, y_n)$.

How is this done? Typically, a language understanding module maps the user utterance $x_k$ (in context) to a formal meaning representation. The agent reasons about this meaning representation to determine its own actions $a_k$. Finally, a response generation module maps these actions or their results (in context) to the agent utterance $y_k$.

The focus of this paper is the response generator. We assume that the formal meaning representation takes the form of an executable program, as is common in the semantic parsing literature—and that the actions are produced by evaluating this program, possibly with side effects. As described by Semantic Machines et al. (2020), the program can be viewed as a dataflow graph in which each node is labeled with a function, constructor, or primitive value, as well as a return value once the node is executed. We aim to implement a response generator that, when applied to an evaluated dataflow graph, satisfies the three properties outlined in §1: description of data and its provenance, guaranteed truthfulness, and declarative specification. In practice, for guidance when developing our generator, we refer to a development set of dialogues annotated with gold-standard dataflow graphs and agent responses.

3 Dataflow Transduction

Given a dataflow graph $G$ (e.g., Fig. 1a) rooted at a node $v_{\text{root}}$ (the return value of the program represented by the dataflow graph), our task is to generate a string that describes $v_{\text{root}}$ and its provenance. To achieve this, we propose a new formal framework for generation based on dataflow transduction. At a high level, the formalism uses declarative rules that describe how to transform a dataflow graph into a small graph-specific grammar (specifically a quasi-synchronous context-free grammar, or QCFG) that defines the space of allowed responses. These rules walk along the graph, introduce new computations (dataflow subgraphs) as needed, and add rules to the grammar.

Formally, a dataflow transducer $S$ is defined by a 4-tuple $(T, \Sigma, R, t_{\text{start}})$ where $T$ is a set of nonterminal types, $\Sigma$ is the set of terminals (word types), $R$ is a set of dataflow transduction rules (see §3.1), and $t_{\text{start}} \in T$ is the start nonterminal. When applied to $G$, the dataflow transducer produces a QCFG. As a side effect, it may extend the graph with new computations. We use $\tilde{G}$ to denote the extended graph.

A QCFG (Smith and Eisner, 2006) is a specialized CFG whose nonterminals include alignments to the nodes $V(\tilde{G})$ of $\tilde{G}$. Where an ordinary CFG might specify ways to generate an NP (noun phrase) or a DATE, a QCFG would specify ways to generate an NP or DATE that describes the result and provenance of $v$, for each appropriately typed node $v \in V(\tilde{G})$. A QCFG resulting from dataflow transduction is a 4-tuple $(T \times V(\tilde{G}), \Sigma, P, (t_{\text{start}}, v_{\text{root}}))$ where $T \times V(\tilde{G})$ is the QCFG’s set of nonterminals and $P$ is its set of productions. A QCFG production has the form $\alpha \rightarrow \beta_1 \beta_2 \cdots \beta_N$ where the left-hand-side $\alpha = (t, v) \in T \times V(\tilde{G})$ is a QCFG nonterminal, and each $\beta_i$ can be either a nonterminal $(t_i, v_i)$ or a terminal in $\Sigma$. The $v_i$ of a right-hand-side nonterminal $\beta_i$ may have appeared in the original $G$, or may have been added to $\tilde{G}$ by the dataflow transducer. These production rules then derive a set of strings just as in any CFG.

3.1 Dataflow Transduction Rules

A dataflow transduction rule is applied to a node $v \in V(\tilde{G})$ (if $v$ has appropriate properties) to cre-

---

1In practice, nonterminal types might correspond to dialogue acts, syntactic categories, semantic categories, etc. This is up to the system designer.
Figure 2: A dataflow transduction rule with head $S$, a body (expressed in Python), and a response template (which queries the dictionary returned by the body).

\[ \text{Head: } S \]
\[ \text{Body: } \]
\[ \text{match computation:} \]
\[ \text{case findEventsOnDate(date):} \]
\[ \text{num = size(computation)} \]
\[ \text{event = head(computation)} \]
\[ \text{return ("num": num, "event": event, "date": date)} \]
\[ \text{Response Template:} \]
\[ I \text{ found } \{ \text{LEX } <\text{num}> \} \text{ event } \{ PP <\text{date}> \}. I \text{ is } \{ \text{EVENT } <\text{event}> \}. \]

3.2 Dataflow Transduction Procedure

Given a dataflow transducer $S$ and a dataflow graph $G\_v$ rooted at node $v\_\text{root}$, we can transduce the graph into a QCFG as follows. The system starts out by creating QCFG productions that can expand the start nonterminal $(t\_\text{start}, v\_\text{root})$. For each transduction rule in $R$ whose head is $t\_\text{start}$, it executes the body, which checks any additional conditions for whether the rule can be applied to $v\_\text{root}$, binds variables, and uses the response template to create a QCFG production. If these productions mention new nonterminals, the system recursively creates further QCFG productions, in the same way, that can expand those nonterminals. As a special case, to expand a nonterminal of the form $(\text{LEX}, v)$, the system creates a QCFG production whose right-hand side gives the value of $v$, as rendered into natural language using a lexicalization function rather than a template; e.g., a value $\text{Integer}(42)$ would be rendered as “42”.

The recursive process continues until productions have been created for every nonterminal that appears in the QCFG. The resulting QCFG compactly represents a combinatorial space of possible responses. It will generally include multiple productions aligned to the same node $v$, created by different dataflow transduction rules.

This mechanism can be used to copy simple values like strings and numbers from the dataflow graph, as well as to create more complex recursive descriptions. Note that (1) transduction rules are selected via their head but also condition on the dataflow graph through their body, and (2) all QCFG nonterminals are grounded in the dataflow.

2 These nodes may already exist in $G\_v$, or may represent new computations that take existing nodes of $G\_v$ as input.

3 This is equivalent to specifying multiple rules with the same head and body, but more concise.
Figure 3: The hybrid response generation approach using dataflow transduction and constrained decoding. Given a computation `nonEmpty(findEventsOnDate(tomorrow()))` for the user utterance “Do I have any meetings tomorrow?”, we first derive QCFG productions by applying the dataflow transducer to the dataflow graph $G$ using the procedure described in §3.2. This procedure also expands the dataflow graph into $G$: for example, the nodes $v3$ and $v4$ were added by the third transducer rule. Then we extract candidate responses from a LM, constrained by the QCFG. The varying descriptions of the date $v2$ and the event $v4$ are permitted because the QCFG offers a choice of productions that can be used to expand the $(PP, v2)$ and $(EVENT, v4)$ nonterminals. (Those productions and the transducer rules that created them are not shown in the figure. The nodes added by those transducer rules and used by those productions are also not shown, except for $v5$.)

Note there may be multiple transduction rules for each QCFG nonterminal $\beta_i$ and the QCFG may admit combinatorially many derivation trees. Each of these derivation trees derives a truthful response. However, since different trees use different rules, the responses may vary in their information content, presentation order, linguistic style, and choice of terminals. The amount of variation can be controlled by the author of the dataflow transducer. In this paper, we use a neural LM with constrained decoding to select a fluent and appropriate response from all these truthful responses, as described in the next section (§4).

4 Constrained Decoding

In this section, we describe how to integrate the formal framework above with a general LM to perform response generation, as illustrated in Fig. 3. Given a derived QCFG of the kind described in §3.2, we will perform constrained decoding as in (Shin et al., 2021; Roy et al., 2022), generating response candidates from a pretrained LM.

The QCFG resulting from dataflow transduction implicitly represents a set of possible derivation trees and the agent responses they yield. As long as translation rules faithfully describe the nodes they apply to, every derivation in this set will correspond to a truthful agent utterance. But these utterances may not always be grammatical or natural. For example, the response template in Fig. 2 may be realized as “I found 2 event on Monday” since the rule body does not check whether the value of num is 1. Similarly, the response template {EVENT ⟨event⟩} starts on {DATE ⟨date⟩}. may be realized as The product meeting on Monday starts on Monday, if the grammar permits identifying events by their dates. With carefully engineered and highly specialized rules (e.g., using extremely fine-grained nonterminal types), it would be possible to ensure that the responses are always fluent and even that there is always a single possible outcome from the top-down search procedure. However, this would usually require much a more complicated set of rules, which creates a bur-
Our proposed approach instead uses a large-scale pretrained LM (preferably fine-tuned) to select among truthful utterances produced by the QCFG. One option is to use the LM to re-rank all strings that can be produced by the QCFG, but that would be very computationally expensive even when that set is finite. Instead, we follow Shin et al. (2021) and Roy et al. (2022), who decode sentences from a given LM under the constraint that they must be valid under a given CFG. In contrast to these prior papers, which used a static CFG, we derive a new CFG each time the dialogue agent needs to generate a response, by applying the dataflow transducer to the current dataflow graph.

The constrained decoding process is a special case of beam search. For each \( \ell = 0, 1, \ldots \), it maintains up to \( K \) prefixes of the same length \( \ell \) and tries to extend each in all legal ways to length \( \ell + 1 \), pruning back to the \( K \) most probable extensions. For each prefix \( y_1 y_2 \ldots y_\ell \) and each terminal symbol \( y_{\ell + 1} \in T \), the extension \( y_1 y_2 \ldots y_{\ell + 1} \) is only legal if it is a prefix of some legal complete response—i.e., some string that is grammatical under the QCFG. This check can be efficiently performed via an incremental context-free parsing algorithm (Earley, 1970) using the parsing state of the prefix \( y_1 y_2 \ldots y_\ell \). In other words, constrained decoding only considers a prefix if it could be extended into at least one legal complete response. Note that the combinatorially many legal responses are never enumerated individually. Rather, the set is compactly represented by the set of QCFG productions.

5 Experiments

To evaluate the proposed approach, we conducted a set of detailed experiments (§5.1–§5.3) on a subset of the SMCalFlow dataset (Semantic Machines et al., 2020), and a brief study (§5.4) applying our approach to the MultiWOZ dataset (Budzianowski et al., 2018).

5.1 Data and Evaluation Metrics

SMCalFlow is a large-scale task-oriented dialogue dataset, in which each user utterance is annotated with a correct dataflow program (i.e., computation) and a “gold” response that would be desirable for the agent to produce.\(^5\) We use the v2.0 release processed by Platanios et al. (2021). We focus on a subset of SMCalFlow involving calendar event queries. This subset contains 8938 training examples and 1041 validation examples. We found that 187 transduction rules, written by some of us in a matter of hours, were sufficient to cover all gold system responses in these examples.\(^6\) We package the annotated examples, transduction rules, and necessary meta information for executing the dataflow programs as a new dataset, SMCalFlow2Text.

Automatic Metrics. For automatic evaluation, we use several reference-based metrics: BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) are computed using GEM-metrics,\(^7\) and BERTScore-F1 is computed using HuggingFace Evaluate.\(^8\) Following the recommendation of Zhang et al. (2020a), we use the re-scaled version of BERTScore, which is easier to interpret. We additionally consider exact match scores, i.e., \( R@K \), which measure whether one of the top \( K \) response candidates exactly matches the reference. Both \( R@1 \) and \( R@5 \) scores are reported. We lowercase all the strings and remove any extra spaces in the predictions and references before computing the evaluation metrics.

Human Evaluation. It is well-known that popular automatic evaluation metrics may not always reflect the true quality of the generated responses (Celikyilmaz et al., 2021). Thus, we further carry out human evaluation on 297 examples randomly sampled from the validation data. Specifically, for each generated response, we collect human judgments on three questions: grammaticality (“has the virtual assistant made any grammar errors?”), relevance (“has the virtual assistant misunderstood the user’s request?”), and truthfulness (“has the virtual assistant provided any incorrect information as judged using the database and timestamp?”). Three judgments are collected.

---

\(^5\)The “gold” responses were generated from an earlier system, but were manually validated by human experts. Like ours, the earlier system also contained rules and constraints.

\(^6\)Some of our rule bodies chose to expand the dataflow graph by calling functions, so we also had to implement those functions. In an end-to-end dialogue system, most of those functions would already have been implemented to support agent actions, not just natural language responses.

\(^7\)https://github.com/GEM-benchmark/GEM-metrics

\(^8\)https://github.com/huggingface/evaluate
for each question, and we report the percentage of examples where “no” is the majority-voted answer. Higher percentages are better. Crowdworkers are recruited from Amazon Mechanical Turk with qualification requirements such as having a work approval rate higher than 80% and having performed a minimum of 100 annotations. They are paid at the rate of $0.15 per judgment. For responses generated by the constrained decoding approach, annotators generally agree with each other on the three questions, i.e., the percentage of examples where all three workers choose the same answer are around 90%, 78% and 76%, respectively. More details are provided in Appendix C.

5.2 Main Results

Table 1 shows our main evaluation results on SMCalFlow2Text. The first baseline we considered is to randomly sample responses from the generated QCFG. The other baseline is unconstrained LM decoding without using dataflow transduction. Model outputs are compared to “gold” agent utterances. For both unconstrained and constrained decoding, the text used to prompt the LM is a string representation of the computation graph (in the format released in SMCalFlow v2.0), followed by its execution result rendered as a JSON string. In both cases, we decode using beam search with beam size $K = 5$. The LM is initialized from CodeT5-base (Wang et al., 2021) and fine-tuned on all training examples. See Appendix D for more details.

As expected, the QCFG random sampling baseline struggles on all the automatic metrics, since dataflow transduction rules are written with an emphasis on truthfulness rather than fluency. This is reflected in the grammaticality score from the human evaluation as well. However, the truthfulness score matches that of the gold responses (92.3%); this baseline rarely generates incorrect responses. Its responses are sometimes generic and omit information that would be relevant to the user—its relevance score is the lowest among all compared approaches—although this behavior contributes to the high truthfulness score.

In contrast, unconstrained decoding LM without dataflow transduction achieves impressive scores on automatic evaluation. Human evaluation also suggests that the generated responses are grammatically correct and relevant to the user’s request in most cases. However, unconstrained decoding scores low on truthfulness, making false statements in about one-fifth of the generated responses. This high rate of factual errors from neural LMs is consistent with findings in prior work (Wiseman et al., 2017; Maynez et al., 2020; Ruanak et al., 2021; Liu et al., 2023; Zhang et al., 2023). It is usually unacceptable in real-world applications.

Compared with unconstrained decoding, our proposed QCFG-constrained decoding achieves significantly better scores on exact match, truthfulness, and even relevance, while maintaining similar scores on BLEU, ROUGE, BERTScore and grammaticality. In particular, human evaluation results indicate that the quality of generated responses is very close to that of the gold responses. Table 2 shows some example predictions. We share some qualitative analysis in Appendix E.

Since even the gold responses did not achieve 100% on human evaluation scores, we manually inspected those problematic examples. There are 4 examples for which the majority-voted answer to the ungrammaticality question is “yes but understandable,” and others are all rated as not containing any grammar errors. For the relevance question, 4 examples are due to arguably bad data and 2 examples receive tied votes. For the truthfulness question, 9 examples are due to arguably bad data.
Table 2: Example predictions from QCFG-constrained decoding and with unconstrained decoding. In all the examples shown, outputs from unconstrained decoding are untruthful to the database due to content hallucination even though the model has access to the correct execution results as part of the prompt. We observe that in a few cases, the constrained model prefers truthful but pragmatically unhelpful omissions like such as “Looks like it” (in Example C) compared to a more specific response.

<table>
<thead>
<tr>
<th>User Utterance</th>
<th>Gold</th>
<th>QCFG-Constrained Decoding</th>
<th>Unconstrained Decoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Do I have any events scheduled on Sunday the 3rd?</td>
<td>No, I didn’t find any event on March 3rd.</td>
<td>No, I didn’t find any event on March 3rd.</td>
<td>No, I didn’t find any event on May 3rd.</td>
</tr>
<tr>
<td>B. Does Pikachu fly tomorrow?</td>
<td>Yes, I found one event matching “flying in” with Pikachu tomorrow. It’s “fly in” from 4:00 to 4:30 PM.</td>
<td>Yes, I found one event matching “flying in” with Pikachu tomorrow.</td>
<td>Yes, I found one event matching “flying in” with Pikachu tomorrow. It’s “Fly in” from 5:00 to 5:30 PM.</td>
</tr>
<tr>
<td>C. Can you find me any events during last March?</td>
<td>Yes, I found one event last March. It was “Dancing in Athenny” on March 17, 2019 from 4:00 to 4:30 PM.</td>
<td>Looks like it.</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: SMCalFlow ablation results, varying the amount of fine-tuning data (groups 1–3) and the context used in the prompt (groups 4–5). X and ✓ on the first column use unconstrained and QCFG-constrained decoding, respectively.

<table>
<thead>
<tr>
<th>BLEU</th>
<th>ROUGE</th>
<th>BERTSc.</th>
<th>R@1</th>
<th>R@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. LM without fine-tuning</td>
<td>✓</td>
<td>.00</td>
<td>.03</td>
<td>−.47</td>
</tr>
<tr>
<td>2. LM fine-tuned on 3% training data</td>
<td>✓</td>
<td>.68</td>
<td>.81</td>
<td>.80</td>
</tr>
<tr>
<td>3. LM fine-tuned on full training data</td>
<td>✓</td>
<td>.77</td>
<td>.87</td>
<td>.87</td>
</tr>
<tr>
<td>4. LM prompted without execution results</td>
<td>✓</td>
<td>.58</td>
<td>.70</td>
<td>.72</td>
</tr>
<tr>
<td>5. LM prompted with user utterance</td>
<td>✓</td>
<td>.77</td>
<td>.87</td>
<td>.87</td>
</tr>
</tbody>
</table>

8 examples are due to to crowd worker mistakes, and 6 examples receive tied votes.

5.3 Ablation Study

We next analyze how the amount of fine-tuning data and the context used in the prompt impact the quality of generated responses. Results are summarized in Table 3.

Impact of fine-tuning: Without fine-tuning the LM, neither unconstrained nor constrained decoding works well. This is likely due to the mismatch between the pre-training tasks and the response generation task. However, after fine-tuning on only a random 3% of the training data, both approaches achieve significantly better scores, with larger gains on QCFG-constrained decoding. This suggests that QCFG-constrained decoding is much more data-efficient in the low-data regime (268 examples). Indeed, QCFG-constrained decoding using 3% of the training data is on par with unconstrained decoding using 100% of the training data, indicating that several expert hours spent on creating dataflow transduction rules hold almost equivalent value to a large volume of training data. While gaps between unconstrained and QCFG-constrained decoding on automated metrics are small in the full-data setting (Table 1), unconstrained decoding still performs poorly on the truthfulness evaluation. Thus, truthfulness failures from unconstrained decoding are not straightforwardly solved by scaling up training data; QCFG-constrained decoding offers an easier path to faithful response generation.

Impact of context: Results in groups 3–5 in Table 3 all use 100% of the training examples to fine-tune the LM. The difference is in the context used in the LM prompt (during both training and inference). For group 3, the text used to prompt the LM is the computation concatenated with the execution result, which is the same setup used in §5.2. For group 4, we omit the execution results.
from the LM prompt (but not from the decoder constraints), whereas for group 5, we add the user utterance (prefixed to the computation). Comparing group 3 and group 4, omitting execution results significantly harms the performance of unconstrained decoding. In contrast, dataflow transduction rules can execute the computation internally, and do not require the LM to condition on it. Comparing group 3 and group 5, adding the user utterance to the LM prompt does not benefit both approaches much.

5.4 Experiments with MultiWOZ Dataset
To demonstrate the general applicability of our approach for response generation, we carry out a brief study on the widely used MultiWOZ 2.1 dataset (Budzianowski et al., 2018; Eric et al., 2020). We automatically convert the agent action annotations to dataflow computations and write 14 transduction rules. For generating responses, we use the predicted agent actions from the MTTOD system (Lee, 2021). Similar to our experiments on SMCalFlow, we fine-tune CodeT5-base on all training examples, using the ground-truth belief state and predicted action as the text used to prompt the LM. For evaluation, we randomly sample 100 examples from the test split, and two authors manually rate the generated responses from our QCFG-constrained decoding system and the MTTOD system. The inter-annotator agreement is 100%. Almost all generated responses are grammatically correct and relevant to the user utterance. To rate truthfulness, we use the predicted actions as the references. Our QCFG-constrained decoding approach produces truthful responses for all 100 examples, whereas only 89 responses from the MTTOD system are truthful with respect to the predicted actions. Among the 11 remaining examples, 7 of them are due to imperfect delexicalization and 4 are due to hallucination.

6 Related Work
One line of response generation research focuses on generating fluent and coherent responses directly from user utterances without any intermediate structured representation. This paradigm is mostly used for chatbots, as in early rule-based systems (Weizenbaum, 1966; Wallace, 2009), neural conversation models (Vinyals and Le, 2015; Shang et al., 2015; Sordoni et al., 2015; Li et al., 2016; Serban et al., 2016), and recent large-scale pretrained LMs like DialoGPT (Zhang et al., 2020b) and GPT-3 (Brown et al., 2020).

Another line focuses on generating text from structured data, with applications beyond dialogue response generation. For example, the WebNLG challenge (Gardent et al., 2017) generates natural language descriptions from relation tuples, and Lebret et al. (2016) generate a biography from a structured “infobox” record. Many recent dialogue response generation tasks adopt dialogue-act-based meaning representations, including the MultiWOZ dataset (Budzianowski et al., 2018), the Schema-Guided dialogue dataset (Rastogi et al., 2020), and the E2E NLG challenge (Dusek et al., 2020). In contrast, our response generation task uses computations as the input, which do not directly encode the dialogue acts of the responses. This is a more challenging task, as the system needs to perform extra reasoning to obtain the derived information. In this sense, our task is similar to the one in CoSQL (Yu et al., 2019) and Logic2Text (Chen et al., 2020).

Constrained decoding techniques for neural LMs have been developed for text generation with different types of constraints (Balakrishnan et al., 2019; Dathathri et al., 2020; Lu et al., 2021, 2022). As §4 noted, we follow Shin et al. (2021) but choose our grammar constraints dynamically for each response.

7 Conclusion
We have described a hybrid approach for building dialogue response generation systems. Our approach introduces a new formalism for transducing a dataflow graph into a QCFG, which is then used in a constrained decoder that intersects the QCFG with a neural LM. This formal framework makes it possible to write rules to precisely and truthfully describe data and its provenance while deferring surface realization decisions to a flexible language model.

This new approach outperforms unconstrained conditional language modeling in both automatic and human evaluations, especially on truthfulness. Moreover, using 3% of the training data, the constrained decoding approach is on par with the unconstrained decoding approach when it uses 100% of the training data, indicating that several expert hours spent on authoring rules hold almost equivalent value to a large volume of training data.
8 Limitations and Future Directions

Authoring transduction rules is relatively easy but may still be labor-intensive for complex domains. Future work might explore (semi-)automatically deriving transduction rules from data, learning to synthesize them from domain specifications, or curating a collection of domain-general transduction rules that can be imported into new domains.

Our experiments in this paper generated text only in English. It would be interesting to apply the framework to datasets in other languages, e.g., GlobalWoZ (Ding et al., 2022). While our framework is intended to be agnostic to the output language, our notation for response templates might need to be slightly extended (along the lines of Appendix A) to be more convenient to use with morphologically complex languages or free-word-order languages. In these settings, presumably, the QCFG should systematically generate many inflections or orderings for the LM to choose among.

To support multilingual dialogue systems, future work could consider automatically translating the response templates into additional languages—perhaps by working backwards from automatic translations of natural language responses that use those templates.

Relatedly, we have only tested the proposed approach on dataflow graphs. Future work could apply the method to generate textual descriptions of other graph-structured inputs, such as graph databases or abstract meaning representations (AMR) graphs.

While QCFG productions were unweighted in this paper, giving them weights would allow the QCFG to express its own preferences about which productions to use for a given input. For example, in a product-of-experts architecture, the probability of a given response \( y \), would be proportional to the LM probability of \( y \) times the weights of all productions used in the QCFG derivation of \( y \) (summed over all such derivations). Beam search (§4) could then be carried out using prefix weights (Opedal et al., 2023). The weights could be trained using gold responses.

Weighting the QCFG raises the possibility that the dataflow transduction rules could encode pragmatic context-dependent policies. For example, a dataflow transduction rule could call a neural network to assess the suitability of applying the rule to a given node in the dataflow graph, and then weight the resulting QCFG production accordingly.

Ethics Statement

Our proposed approach strongly outperforms a purely neural model at truthfully describing the result of a computation and its provenance. However, our approach can still make pragmatically unhelpful omissions, making it potentially risky to deploy in some scenarios. Additionally, we leverage pre-trained neural language models such as CodeT5, and as such, we acknowledge that our approach might inherit some biases present in these pre-trained models.

Acknowledgements

We would like to thank Ben Van Durme, Baolin Peng, Subhro Roy, Richard Shin, and Patrick Xia for valuable discussions and feedback. We also thank the anonymous reviewers for their insightful comments and suggestions.

References


A Alternatives in Response Templates

A dataflow transduction rule can be equipped with multiple templates, and our template format also allows choices within a single template. Specifically, our implementation allows the use of vertical bar to encode alternatives within a template, e.g., “I {{ didn’t find any | found no }} {{ matching events | events matching {LEX [subject]} }} on your calendar.” During dataflow transduction, this template is automatically converted into a small system of QCFG productions, i.e., introducing new nonterminals for the alternations.

B Dataflow Transduction Rule Details

In our experiments, there are 9 head types (including the START symbol) for the 187 transduction rules for SMCalFlow2Text, and 3 head types for the 14 transduction rules for MultiWOZ. Our framework is agnostic to the nonterminal types (see footnote 1). We mainly used syntactic categories like NP, PP, DT, VB, UH (interjection), Copula, etc. One potential challenge is that the domain developers may need to have some linguistic knowledge about the syntactic categories. Alternatively, they could use semantic categories.

The complete set of rules for SMCalFlow2Text is available in our released Python code. The 187 transduction rules cover the 8938 and 1041 examples from the training and validation set in the original SMCalFlow data, i.e., the gold agent responses can all be produced from the transduction rules. The authors who wrote the rules were able to look at both the training and validation examples. The remaining training and validation examples in the original SMCalFlow dataset are not covered by these rules.

Below we explain some examples of dataflow transduction rules.

The rule head PP suggests that the computation is described as a preposition phrase. The body simply checks whether the computation being described is a call to the function FullMonthofPreviousMonth and extracts the argument month. The response template lexicalizes the function call as “last” and defers describing the month to appropriate NP rules such as the one below.

For this rule, its body checks the value of the computation rather than its structure. Since the response template has no nonterminal, the body does not return any variable binding. Note returning an empty dictionary is different from returning None (which is the default return value in Python), as the latter indicates that the rule cannot be applied.

C Human Evaluation Details

A screenshot of the MTurk interface for human evaluation is shown in Fig. A1. Table A1 shows the percentages of examples where all three workers choose the same answer for individual systems. It can be observed that the gold responses receive the highest agreements on all three questions. The
QCFG-constrained decoding has slightly higher agreements than the unconstrained decoding. The QCFG random sampling receives a significantly lower agreement on “Grammatical,” probably because this approach may produce ungrammatical responses but people may not agree on whether these are understandable.

D Model Configurations

For SMCalFlow, we fine-tune the CodeT5 model for a fixed number of epochs (=10). For MultiWOZ, we fine-tune the model for at most 10 epochs and do early stopping based on the loss on the development set. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, using a linear learning rate scheduler with an initial learning rate of $5 \times 10^{-5}$. For decoding, we always use a fixed beam size of 5.

The CodeT5-base models used in our experiments have 220 million parameters. We used machines with 32GB V100 GPUs for model fine-tuning while the decoding experiments were carried out on CPU-only machines.

For SMCalFlow experiments, the input sequence to the LM is the string representation of the computation in the lispress format followed by its execution result rendered as a JSON string, e.g., “Plan: (Yield (Event.start ( . . . )) Result: {“type”: “DateTime”, “value”: … } <s>),” where the last token is a special token to separate the input and the output. For the ablative study (group 5) in §5.3, the user utterance is prefixed to the sequence, e.g., “User: When do I have thee oil change on my car scheduled for? Plan: … Result: … <s>”.

Figure A1: A screenshot of the MTurk interface for human evaluation.
For MultiWOZ experiments, the computation is rendered as a raw JSON string that encodes the ground-truth belief state and the predicted system act. There is no execution result for these computations.

### E Qualitative Analysis

We looked at 100 randomly selected examples from the experiments on SMCalFlow from §5.2, and compared the generated responses from both unconstrained decoding and QCFG-constrained decoding with the human-annotated gold responses provided by the dataset. We summarize the differences between the generated and gold responses in Table A2, using the following categories:

**Untruth** The system reports incorrect information.

**Omission** The system fails to mention information mentioned in the gold response.

**Addition** The system mentions additional (correct) information that is not mentioned in the gold response.

**Minor Difference** The system uses a different phrasing than the gold response that nonetheless has the same information and fluency.

**Disfluency** The system output is disfluent.

**Annotation Error** The system output is acceptable but the gold annotation contains a fluency or factuality error.

For unconstrained decoding, 57 out of 100 responses differ from the gold responses, whereas for QCFG-constrained decoding, only 51 of 100 responses differ. This result is consistent with the R@1 column of Table 1 (mismatch rates of 53% and 44% respectively on the full validation set).

As expected, the most noticeable difference is in the number of Untruths. The QCFG-constrained system produced no Untruths. The unconstrained system produced 19%, close to the 18% rate found in the human evaluations in Table 1. We show some examples of Untruths in Table 2.

Conversely, the QCFG-constrained system produces substantially more Omissions than the unconstrained system. Of the 11 omissions produced by the constrained system, 3 are identical to the unconstrained output while 7 are on inputs for which the unconstrained output produced Untruth. In other words, our system successfully removed the 19 Untruths by the system, but in 7 of those cases, it produced a shorter (but still factually correct) input than the preferred gold annotation for that example. We also note that the gold dataset is not consistent in how much information is included in the responses – short answers like “Looks like it” in Example C from Table 2 are present in the gold annotations on examples similar to Example C. Furthermore, both systems produce more Additions than Omissions, indicating that there is not a systematic bias towards shorter answers overall. In future work, the model could be made to select more descriptive responses by adding a brevity penalty in the decoder or by weighting the QCFG productions, so that responses are scored not only by the LM but also by the QCFG.

### F Dataset License

The SMCalFlow dataset is distributed under the CC BY-SA 4.0 license. To the best of the authors’
knowledge, the MultiWOZ datasets were released under MIT license as shown in https://github.com/budzianowski/multiwoz. Our experiments follow the intended use of these datasets, which is to advance research in dialogue systems.