Code: https://github.com/microsoft/dataflow2text
Preprint: https://arxiv.org/abs/2209.07800

# 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

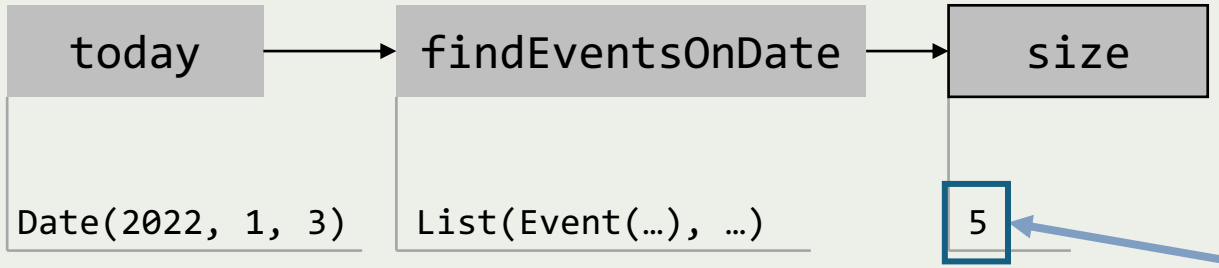
Microsoft Semantic Machines

*: Equal Contribution

Turn 1

User: *How many events are on my calendar today?*

```
size(findEventsOnDate(today()))
```




Agent: *You have three events.* ✗

Agent: *You have five events.* ✓

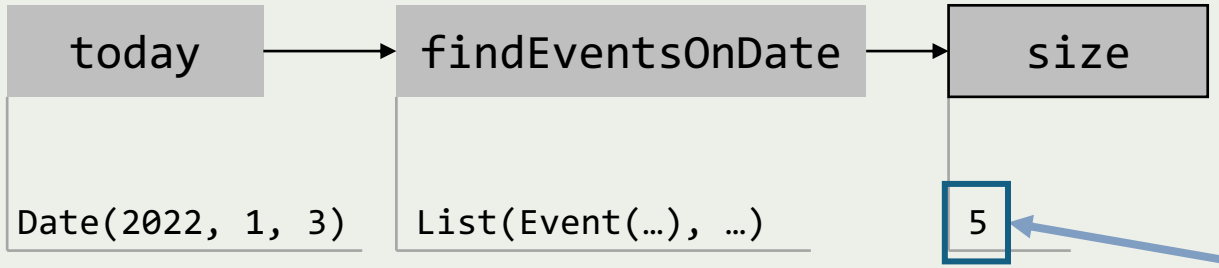
Computation
(Semantic Machines et al., 2020. *"Task-Oriented Dialogue as Dataflow Synthesis"*.)

The response should be ***truthful*** about the execution result.

Turn 1

User: *How many events are on my calendar today?*

```
size(findEventsOnDate(today()))
```




Agent: *You have three events.* ✗

Agent: *You have five events.* ✓

Computation
(Semantic Machines et al., 2020. *"Task-Oriented Dialogue as Dataflow Synthesis"*.)

The response should be ***truthful*** about the execution result.

Turn 2

User: *Can you schedule a meeting with Sarah Smith?*

```
createEvent(attendee=queryPerson(name="Tara Smith"))
```

Agent: *OK. I've booked it.* ✗

Agent: *OK. I've booked a meeting with Tara Smith at 2pm today.* ✓

The response should describe more of the computation graph than just the result. It should confirm what the system actually did.

# Introduction

- Two predominant paradigms for dialogue response generation

**Neural Language Modeling**

- Produce fluent, coherent, diverse responses.
- Can leverage pre-trained large language models (e.g., GPT-3, ChatGPT).
- Issues:
  - Produce unfaithful and unsafe responses.
  - Difficult to control.

**Rule-Based Generation**

- Easy to control (by modifying rules).
- Safe for production (can only produce responses allowed by rules).
- Issues:
  - Hard to maintain for complex domains.
  - Require extensive domain knowledge.

How to combine the strength of both?

# A Hybrid Approach for Response Generation

Dataflow Transducer

+

Language Model

(Dataflow Transduction)
Transduce a computation graph into a context-free grammar.

(Constrained Decoding)
Generate responses constrained by the context-free grammar.

The context-free grammar (CFG) defines the space of all responses allowed for the given computation.

These responses are *truthful* but not always *grammatical* or *natural*.

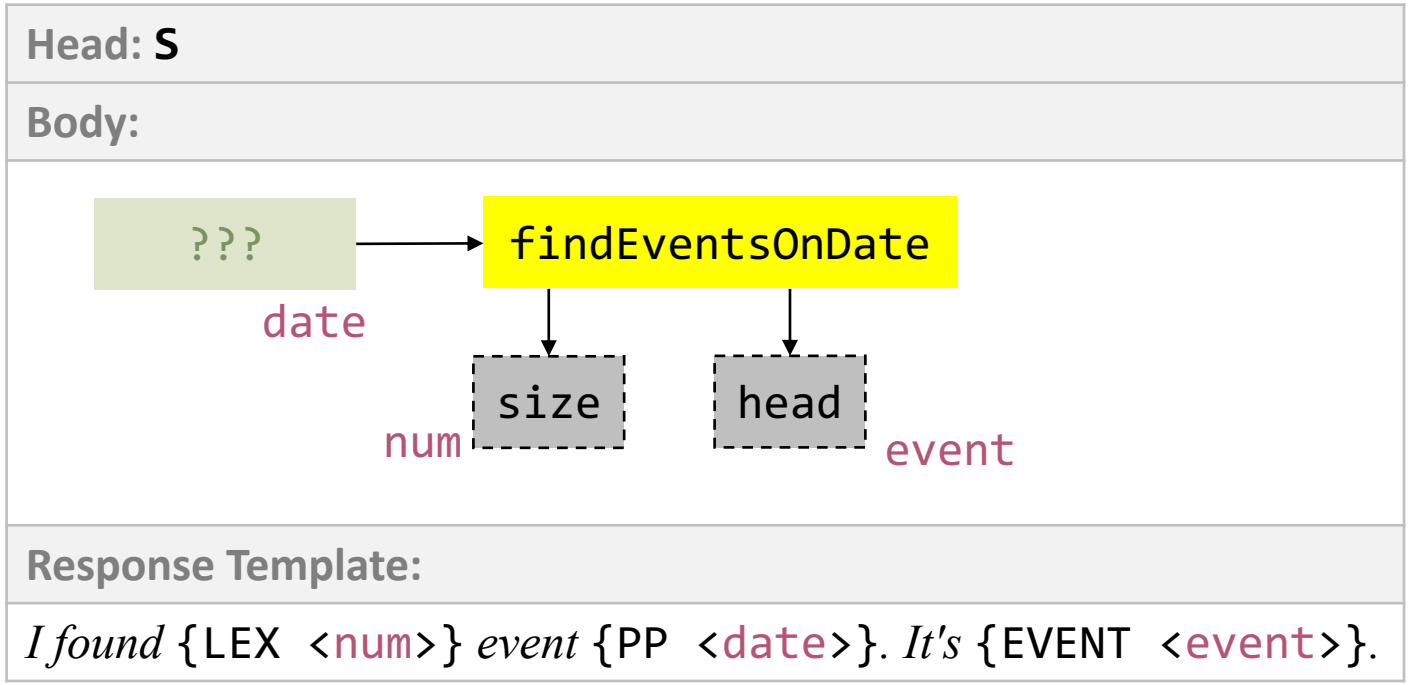
✓ *I found 1 event on Thursday. It's "Show and Tell".*
✗ *I found 1 events on Thursday.*
✗ *The "Show and Tell" meeting on Thursday starts on Thursday.*

Hybrid generation has a long history in NLP, dating back to Knight & Hatzivassiloglou (1995) and Langkilde & Knight (1998).

# Dataflow Transduction Rule

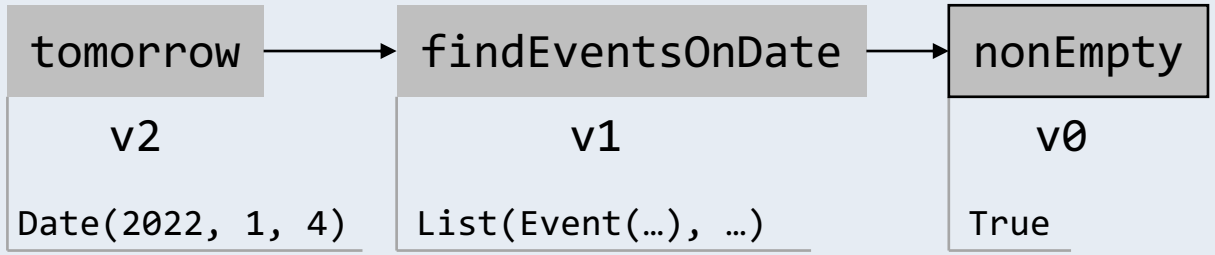
- Applied to a computation node to create a QCFG production

QCFG is a special type of CFG (more details in our paper).



User: *Do I have any meetings tomorrow?*

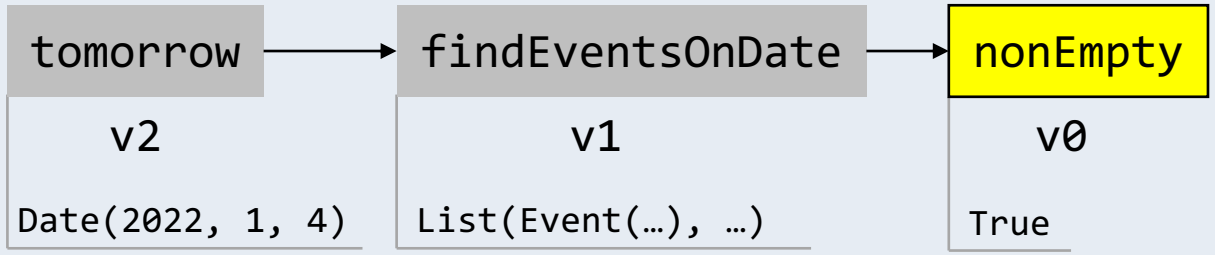
```
nonEmpty(findEventsOnDate(tomorrow()))
```




Agent: *Yes , I found 1 event. It's "Show and Tell" from 11:00 am to 11:30 am on Thursday.*

User: *Do I have any meetings tomorrow?*

```
nonEmpty(findEventsOnDate(tomorrow()))
```



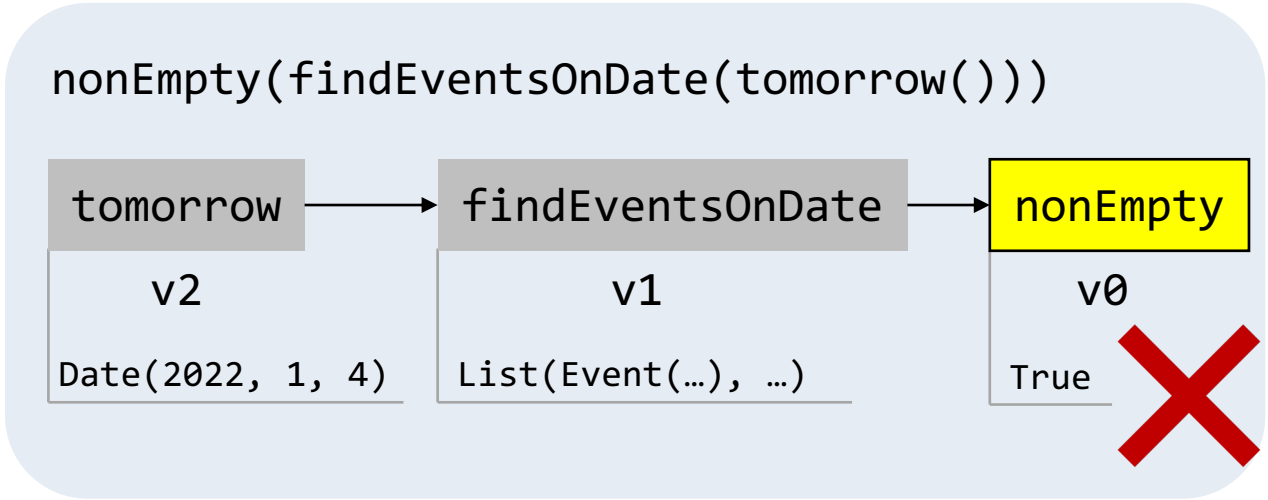

## QCFG Productions

(S, v0) → (Y/N, v0) , (S, v1)

| Head: **S** |
| --- |
| Body: |
| ??? → nonEmpty<br>query result |
| Response Template: |
| {Y/N <result>}, {S <query>}. |

**Agent:** Yes, I found 1 event. It's "Show and Tell" from 11:00 am to 11:30 am on Thursday.

User: *Do I have any meetings tomorrow?*




## QCFG Productions

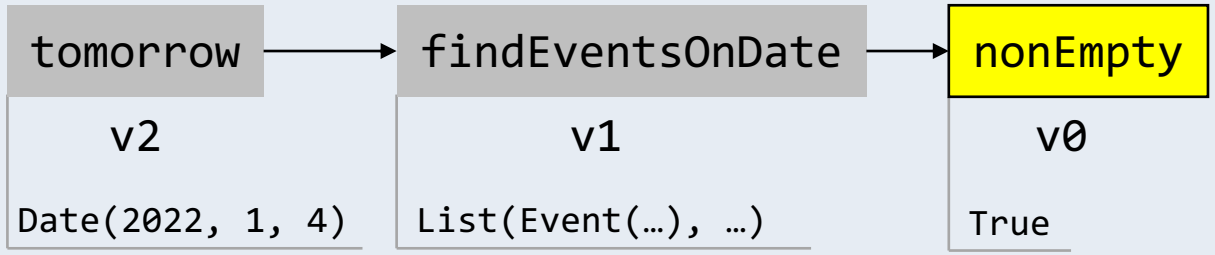
```
(S, v0) → (Y/N, v0) , (S, v1)
(Y/N, v0) →
```

| Head: **Y/N** |
| --- |
| Body: |
| ??? <br> False |
| Response Template: |
| *No* |

Agent: *Yes, I found 1 event. It's "Show and Tell" from 11:00 am to 11:30 am on Thursday.*

User: *Do I have any meetings tomorrow?*

```
nonEmpty(findEventsOnDate(tomorrow()))
```




## QCFG Productions

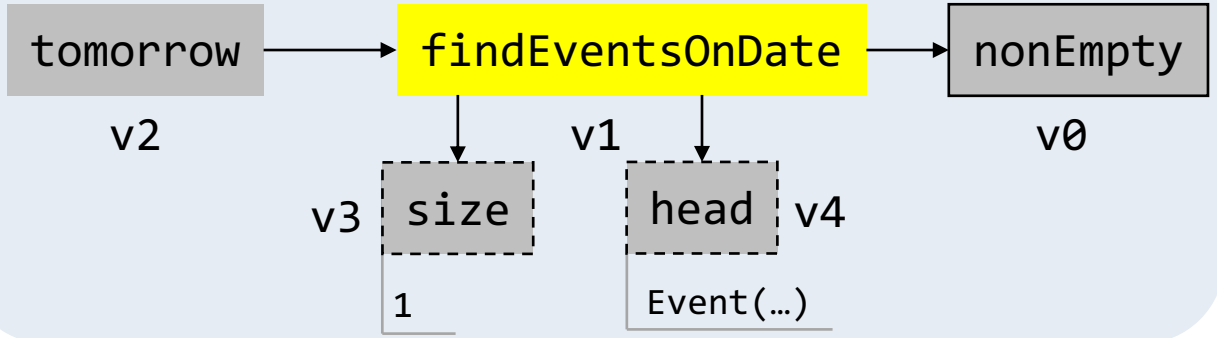
```
(S, v0) → (Y/N, v0) , (S, v1)
(Y/N, v0) → Yes
```

| Head: **Y/N** |
| --- |
| **Body:** |
| ??? <br> True |
| **Response Template:** |
| *Yes* |

Agent: *Yes, I found 1 event. It's "Show and Tell" from 11:00 am to 11:30 am on Thursday.*

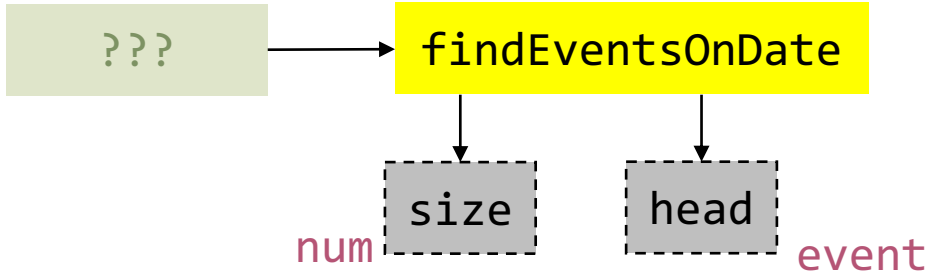
User: *Do I have any meetings tomorrow?*

```
nonEmpty(findEventsOnDate(tomorrow()))
```




**Head:** **S**

**Body:**




**Response Template:**

*I found* {LEX <num>} *event. It's* {EVENT <event>}.

**Agent:** *Yes, I found* [illegible] *event. It's* [illegible] *"Show and Tell" from 11:00 am to 11:30 am on Thursday.*

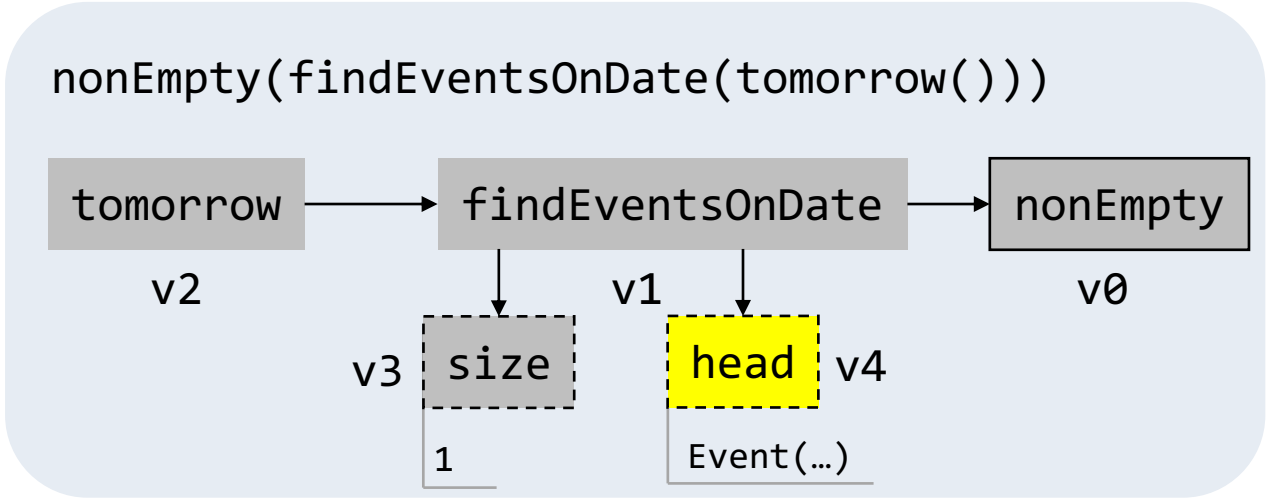
## QCFG Productions

(S, v0) → (Y/N, v0) , (S, v1)

(Y/N, v0) → *Yes*

(S, v1) → *I found* (LEX, v3) *event. It's* (EVENT, v4).

User: *Do I have any meetings tomorrow?*




## QCFG Productions

(S, v0) → (Y/N, v0) , (S, v1)

(Y/N, v0) → *Yes*

(S, v1) → *I found* (LEX, v3) *event .* *It's* (EVENT, v4).

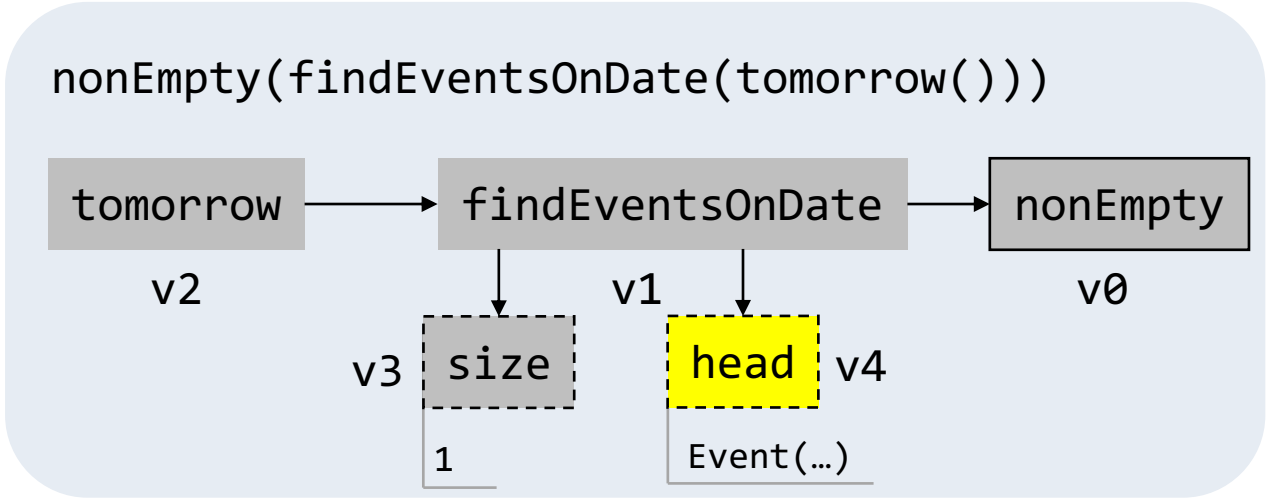
(LEX, v3) → *1*

(EVENT, v4) → ...

- Termination Condition: All nonterminals are expanded.

Agent: *Yes, I found 1 event. It's "Show and Tell" from 11:00 am to 11:30 am on Thursday.*

User: *Do I have any meetings tomorrow?*




## QCFG Productions

(S, v0) → (Y/N, v0) , (S, v1)

(Y/N, v0) → *Yes*

(S, v1) → *I found* (LEX, v3) *event* . *It's* (EVENT, v4).

(LEX, v3) → *1*

(EVENT, v4) → ...

.

.

.

- Termination Condition: All nonterminals are expanded.
- There may be multiple applicable transduction rules for each QCFG nonterminal. The resulting QCFG represents combinatorially many truthful responses.
- We intersect the QCFG with a neural LM to select a fluent and appropriate response from these truthful responses.

...

*(0.8) I found 1 event. It's "Show and Tell" from 11:00 am to 11:30 am.*

*(0.6) I found 1 event. It's "Show and Tell".*

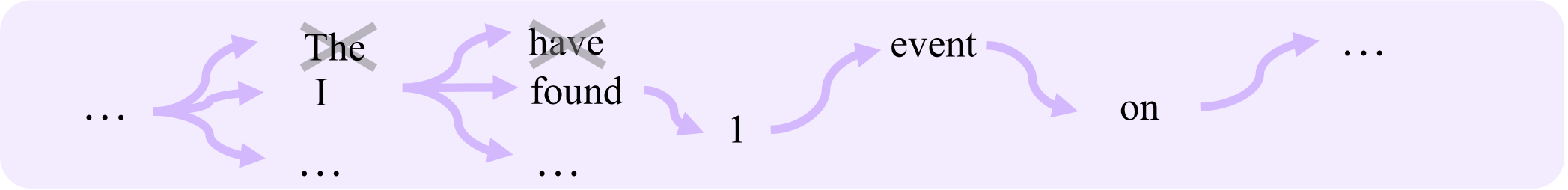
*(0.2) I found 1 event on Thursday starts on Thursday.*

...

Agent: *Yes, I found 1 event. It's "Show and Tell" from 11:00 am to 11:30 am on Thursday.*

# Constrained Decoding

- Generate response candidates from a neural LM (pre-trained and preferably fine-tuned), constrained by the QCFG.




Shin et al., 2020. *"Constrained Language Models Yield Few-Shot Semantic Parsers".*

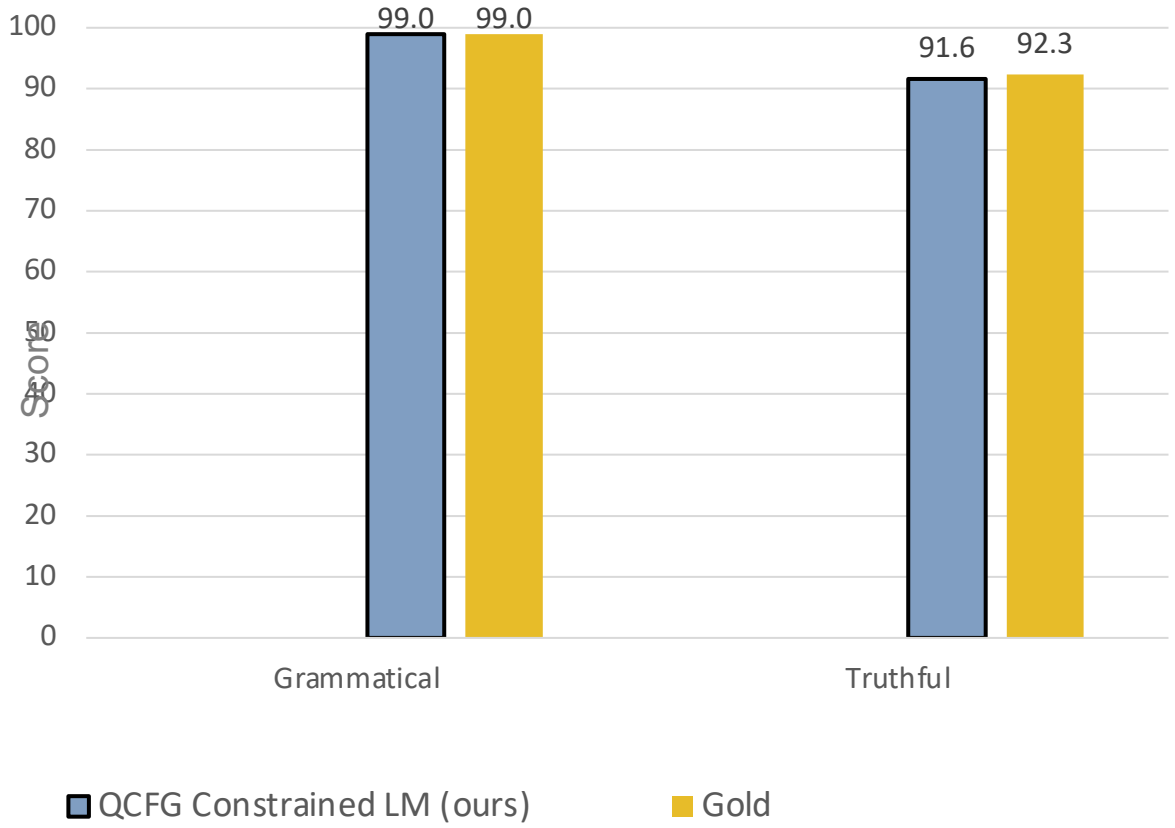
- Can be efficiently performed via an incremental context-free parsing algorithm (Earley, 1970) using the parsing state of the prefix.

# Experiments with SMCalFlow2Text

- Experiments with a subset of SMCalFlow[1] involving calendar event queries
  - 8938 training instances, 1041 test instances
  - Manually authored 187 transduction rules
- CodeT5[2] models fine-tuned on the train split
  - Input is computation graph with execution results
  - Output is agent response

[1] Semantic Machines et al., 2020. *"Task-Oriented Dialogue as Dataflow Synthesis"*.
[2] Wang et al., 2021. *"CodeT5: Identifier-Aware Unified Pretrained Encoder-Decoder Models for Code Understanding and Generation"*.

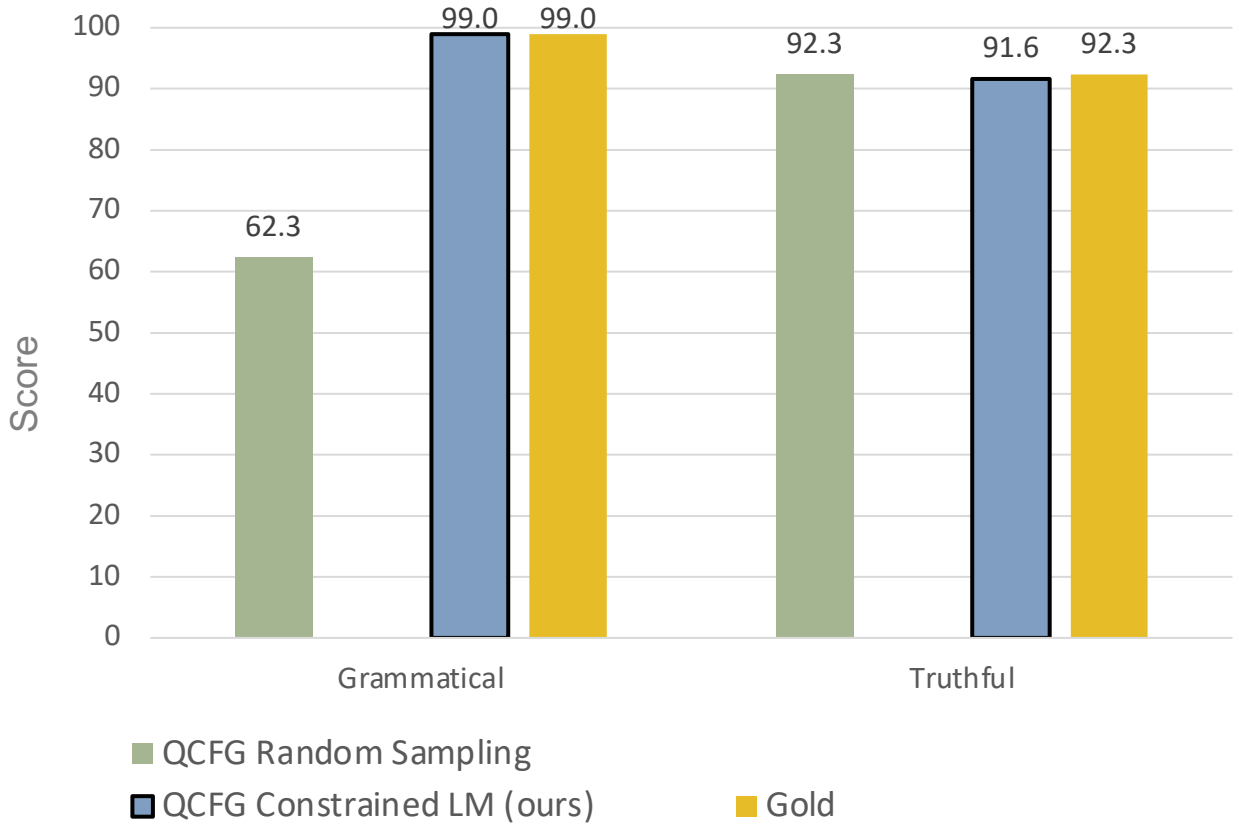
# Human Evaluation on SMCalFlow2Text




Human Evaluation on predictions for 297 randomly sampled test examples.

- Gold outputs score very high on grammatical correctness and truthfulness as expected.
- Constrained decoding from a fine-tuned model produces grammatically correct and truthful responses (very close to gold references).

# Human Evaluation on SMCalFlow2Text

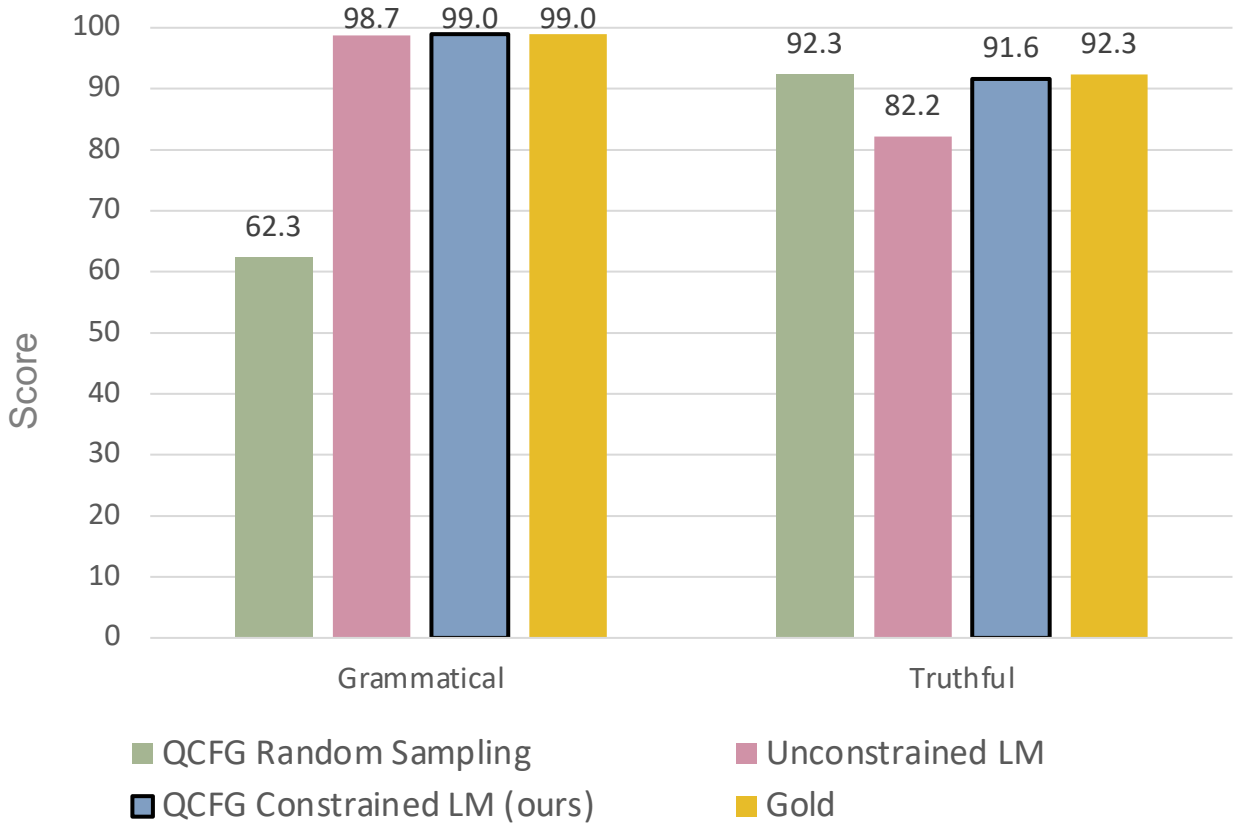



Baseline 1: Random sampling from the QCFG can produce ungrammatical and non-fluent responses

=> Shows the importance of having LM in our approach

*I found 1 events on Thursday.*
*The "Show and Tell" meeting on Thursday starts on Thursday.*

# Human Evaluation on SMCalFlow2Text




Baseline 2: Decoding from the fine-tuned neural model (conditioned on the computation graph and execution results, but *without* constraining using dataflow transduction) can produce grammatically fluent responses, but they score low on truthfulness.

=> Shows the importance of QCFG constraints in our approach

# Qualitative Examples

| | User Utterance | Gold | QCFG-Constrained Decoding | Unconstrained Decoding |
|---|---|---|---|---|
| A. | *Do I have any events scheduled on Sunday the 3rd?* | *No, I didn't find any event on March 3rd.* | *No, I didn't find any event on March 3rd.* | *No, I didn't find any event on May 3rd.* |
| B. | *Does Pikachu fly in tomorrow?* | *Yes, I found one event matching "flying in" with Pikachu tomorrow. It's "fly in" from 4:00 to 4:30 PM.* | *Yes, I found one event matching "flying in" with Pikachu tomorrow. It's from 4:00 to 4:30 PM.* | *Yes, I found one event matching "flying in" with Pikachu tomorrow. It's "Fly in" from 5:00 to 5:30 PM.* |

# Conclusion and Future Directions

- A hybrid approach for building dialogue response generation systems.
  - Developers can write transduction rules to *truthfully* describe computations.
    - multiple rules ⟶ many choices of surface realization (nondeterminism)
  - Surface realization decisions are deferred to a flexible language model.
- Future Directions
  - Use better ways to obtain dataflow transduction rules, e.g., automatically derive from data or synthesize from domain specifications.
  - Support multi-lingual dialogue systems.
  - Generate textual descriptions of other graph-structured inputs, e.g., graph databases or AMR graphs.
  - Weight the QCFG productions to encode pragmatic policies.

# Thank you!


https://semanticmachines.com