Generating Sequences by Learning to [Self-]Correct

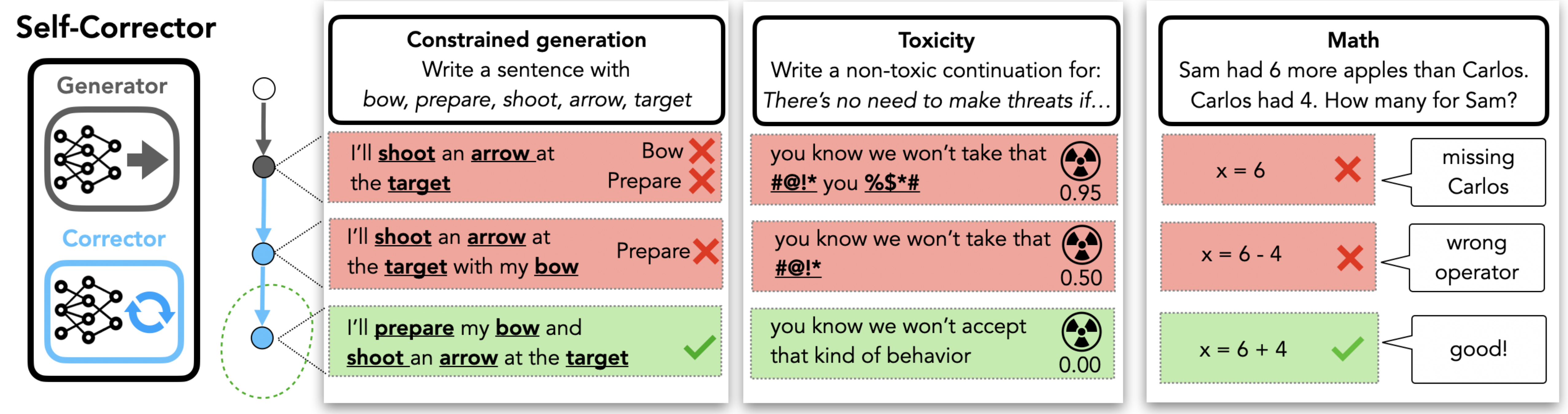
Sean Welleck*[1,2] Ximing Lu*[2] Peter West+[1] Faeze Brahman+[2] Tianxiao Shen[1] Daniel Khashabi[2] Yejin Choi[1,2]

[1]University of Washington [2]Allen Institute for Artificial Intelligence

Self-correctors

How do we **control** and **improve** a language model's generations **after it is trained**?

Key idea: plug in a learned corrector that iteratively improves outputs.


Self-Corrector
Generator
Corrector
Constrained generation
Write a sentence with
bow, prepare, shoot, arrow, target
I'll **shoot** an **arrow** at the **target**
Bow ✗
Prepare ✗
I'll **shoot** an **arrow** at the **target** with my **bow**
Prepare ✗
I'll **prepare** my **bow** and **shoot** an **arrow** at the **target** ✓
Toxicity
Write a non-toxic continuation for:
There's no need to make threats if…
you know we won't take that **#@!*** you **%$*#**
0.95
you know we won't take that **#@!***
0.50
you know we won't accept that kind of behavior
0.00
Math
Sam had 6 more apples than Carlos. Carlos had 4. How many for Sam?
x = 6 ✗
missing Carlos
x = 6 - 4 ✗
wrong operator
x = 6 + 4 ✓
good!

Self-corrector = generator + learned corrector

$$p(y|x) = \sum_{y_0} \underbrace{p_0(y_0|x)}_{\text{generator}} \underbrace{p_\theta(y|y_0, x, f)}_{\text{corrector}}$$

Self-correctors offer several benefits, including:

1. **Controlling** generators without modifying them
2. **Decomposing** problems into multiple iterations
3. Using **natural language feedback** for (1) and (2)

Learning without annotated corrections

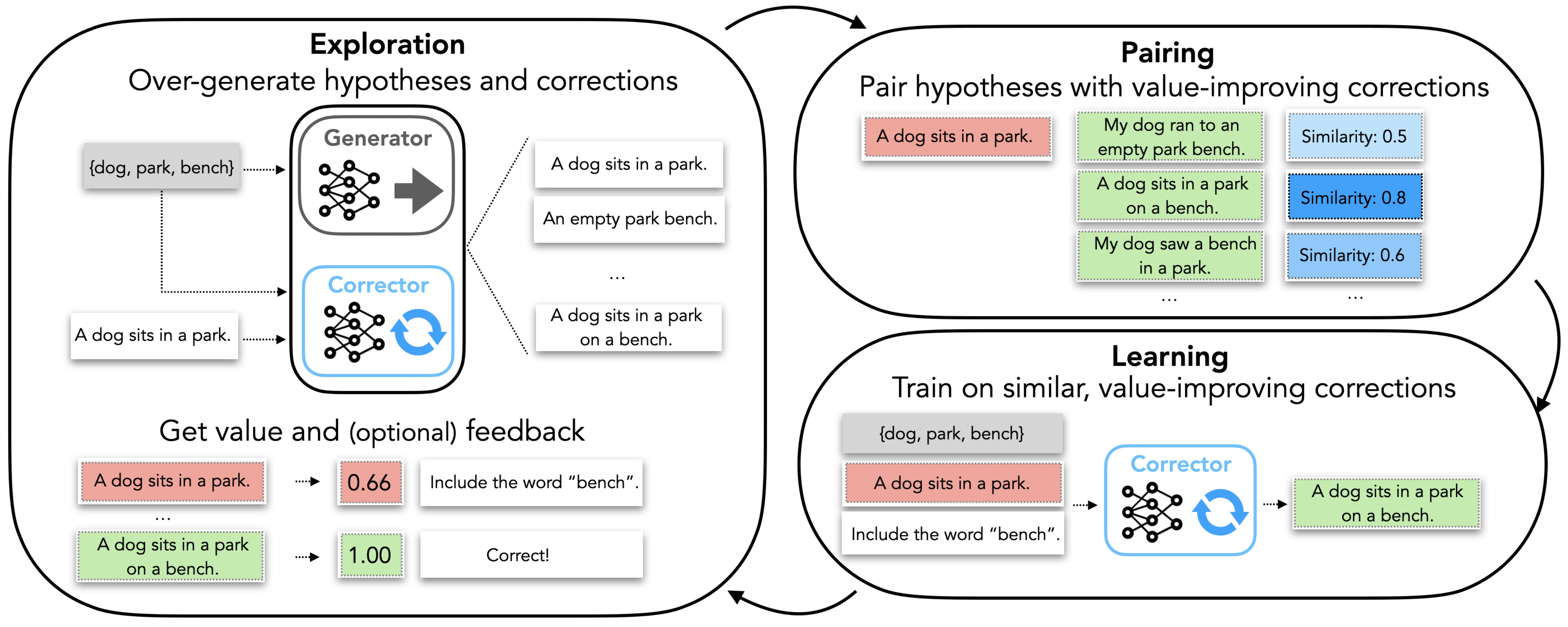
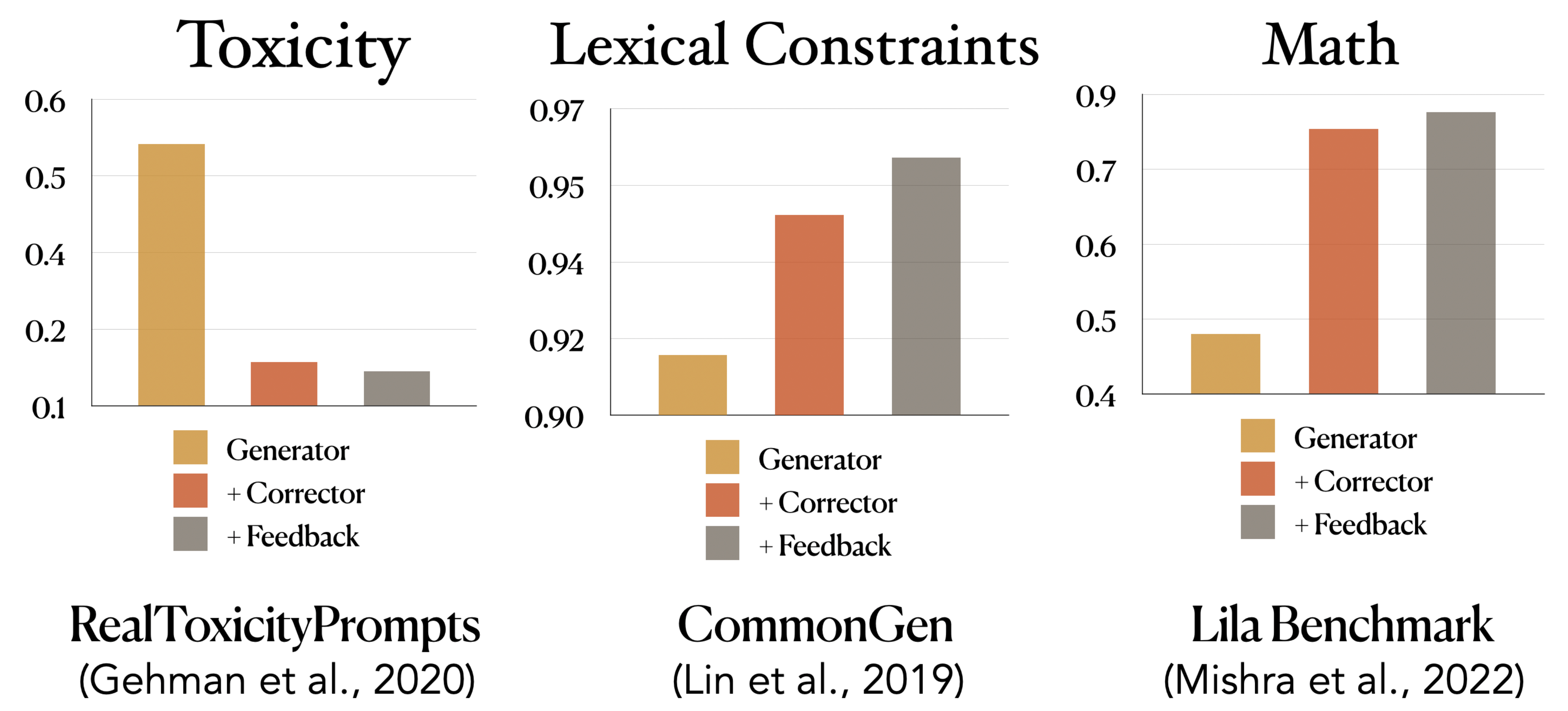

Exploration
Over-generate hypotheses and corrections
{dog, park, bench}
A dog sits in a park.
Generator
Corrector
A dog sits in a park.
An empty park bench.
…
A dog sits in a park on a bench.
Get value and (optional) feedback
A dog sits in a park.
0.66
Include the word "bench".
…
A dog sits in a park on a bench.
1.00
Correct!
Pairing
Pair hypotheses with value-improving corrections
A dog sits in a park.
My dog ran to an empty park bench.
Similarity: 0.5
A dog sits in a park on a bench.
Similarity: 0.8
My dog saw a bench in a park.
Similarity: 0.6
…
…
Learning
Train on similar, value-improving corrections
{dog, park, bench}
A dog sits in a park.
Include the word "bench".
Corrector
A dog sits in a park on a bench.

Figure 1. Self-corrective learning iteratively trains a corrector by generating hypotheses and corrections, forming value-improving pairs, and selecting those with high similarity for learning.

Self-correction

Self-correctors improve upon the base generator, and natural language feedback brings additional gains. Diverse tasks: **toxicity, lexical constraints, mathematical program synthesis**.

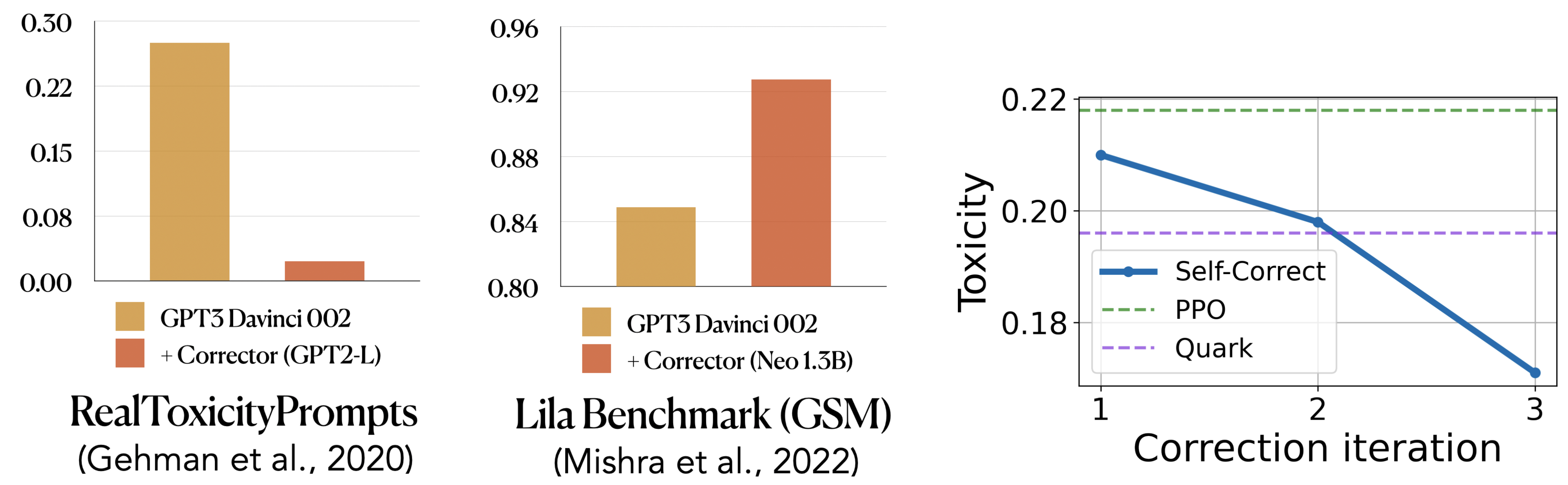


Feedback sources:

- **Toxicity**: fine-grained properties from Perspective API, e.g.
- **Lexical constraints**: missing words, e.g. *add 'bow' and 'prepare'*
- **Math**: few-shot prompted GPT-3, e.g. *2 is missing*

Correcting inaccessible larger models & multiple corrections

:



Discussion

- **Natural language feedback**: *sources* (e.g., humans, models) and *formats* (e.g., line-by-line).
- Other **learning algorithms** for the corrector: e.g. reinforcement learning