Reinforcement Learning via Self-Distillation

Jonas Hübotter, Frederike Lübeck*, Lejs Behric*, Anton Baumann*,
Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld,
Thomas Kleine Buening, Carlos Guestrin, Andreas Krause

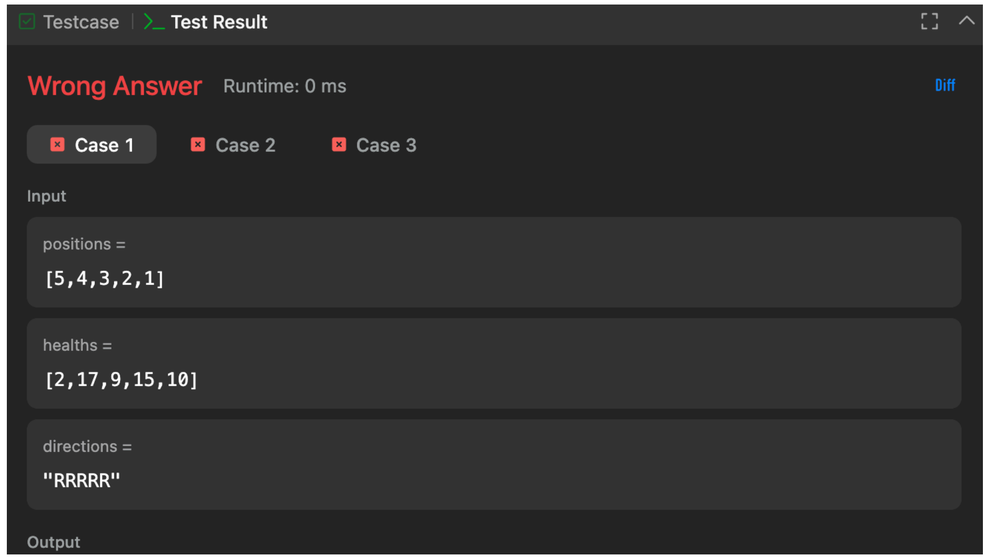
ETH zürich
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
Stanford University

The feedback bottleneck in RL

How do you improve at coding?

The feedback bottleneck in RL

How do you improve at coding?

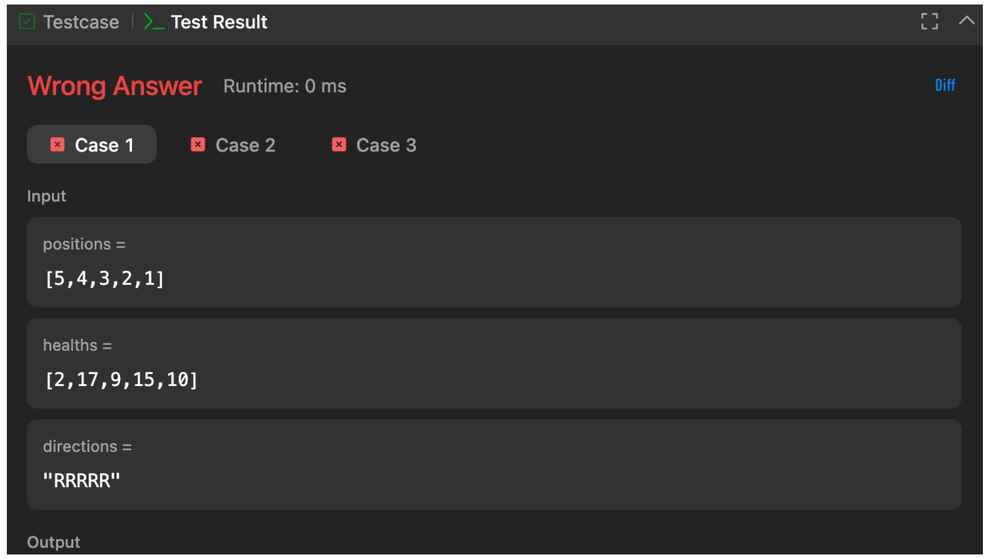


Leetcode

Run unit tests, debug runtime errors, profiling, ...

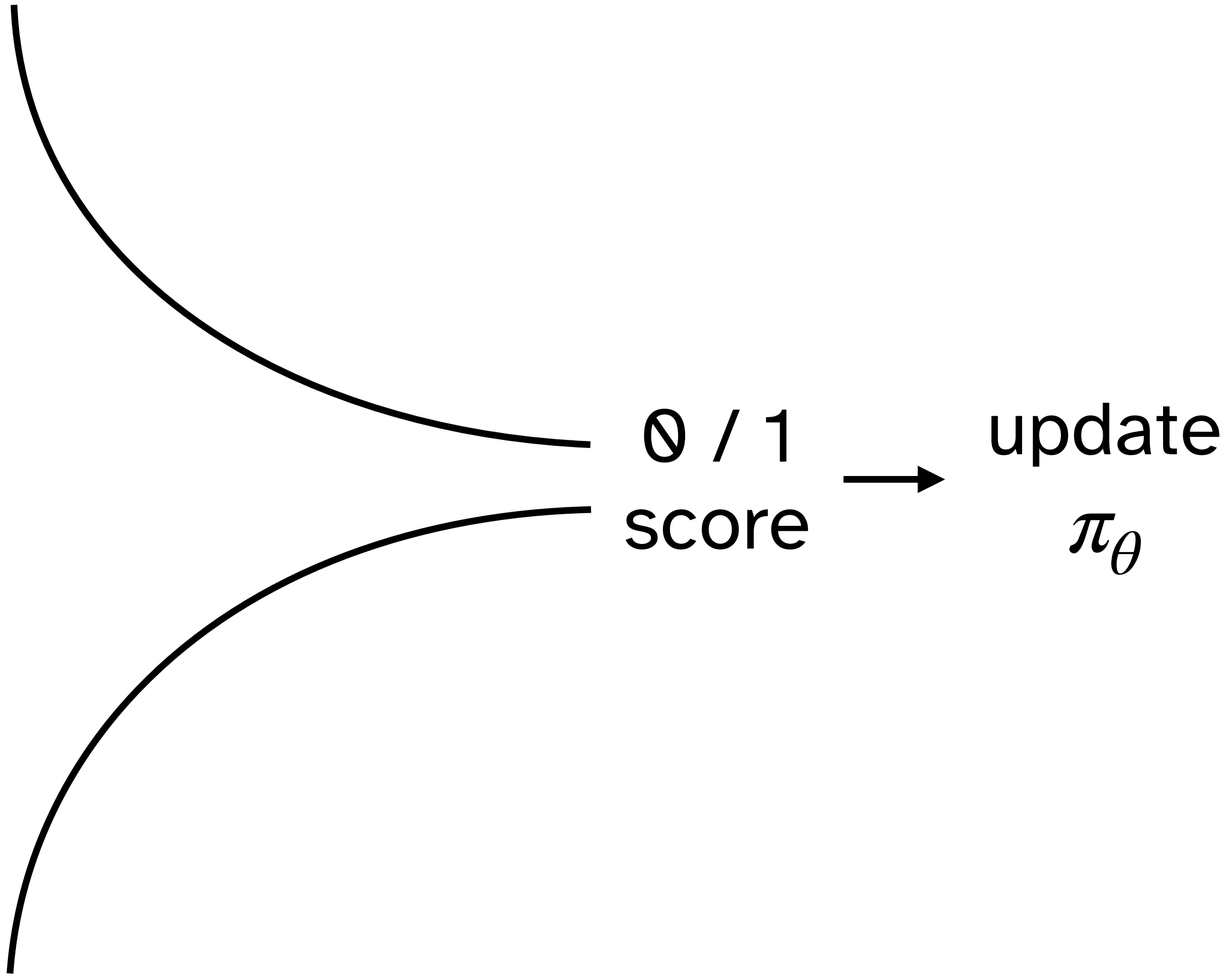
The feedback bottleneck in RL

How do you improve at coding?



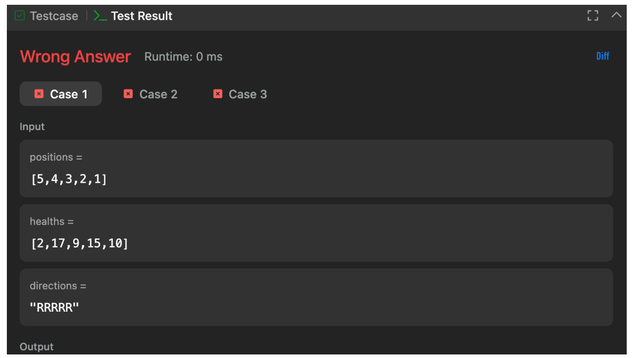
Leetcode

Run unit tests, debug runtime errors, profiling, ...



The feedback bottleneck in RL

How do you improve at coding?



Leetcode

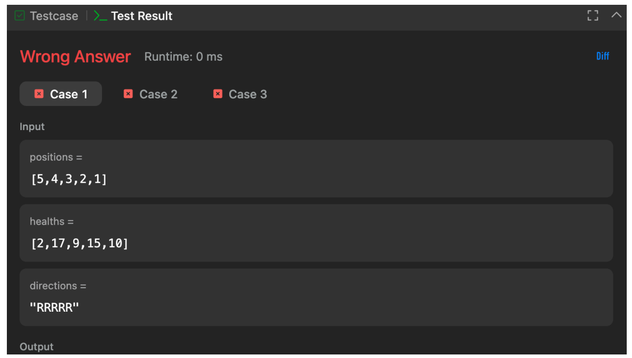
Run unit tests, debug runtime errors, profiling, ...

0 / 1 score → update π_θ

Vast data ← Sparse signal

The feedback bottleneck in RL

How do you improve at coding?



Leetcode

Run unit tests, debug runtime errors, profiling, ...

0 / 1 score → update π_θ

Vast data ← Sparse signal

Scaling the opportunity for post-training

Question:

Write a python function that returns all numbers from 1 to n. Answer briefly.

Answer $y \sim \pi_\theta(\cdot \mid x)$:

```python
def numbers_up_to_n(n):
    return list(range(1, n + 1))
```

incorrect!

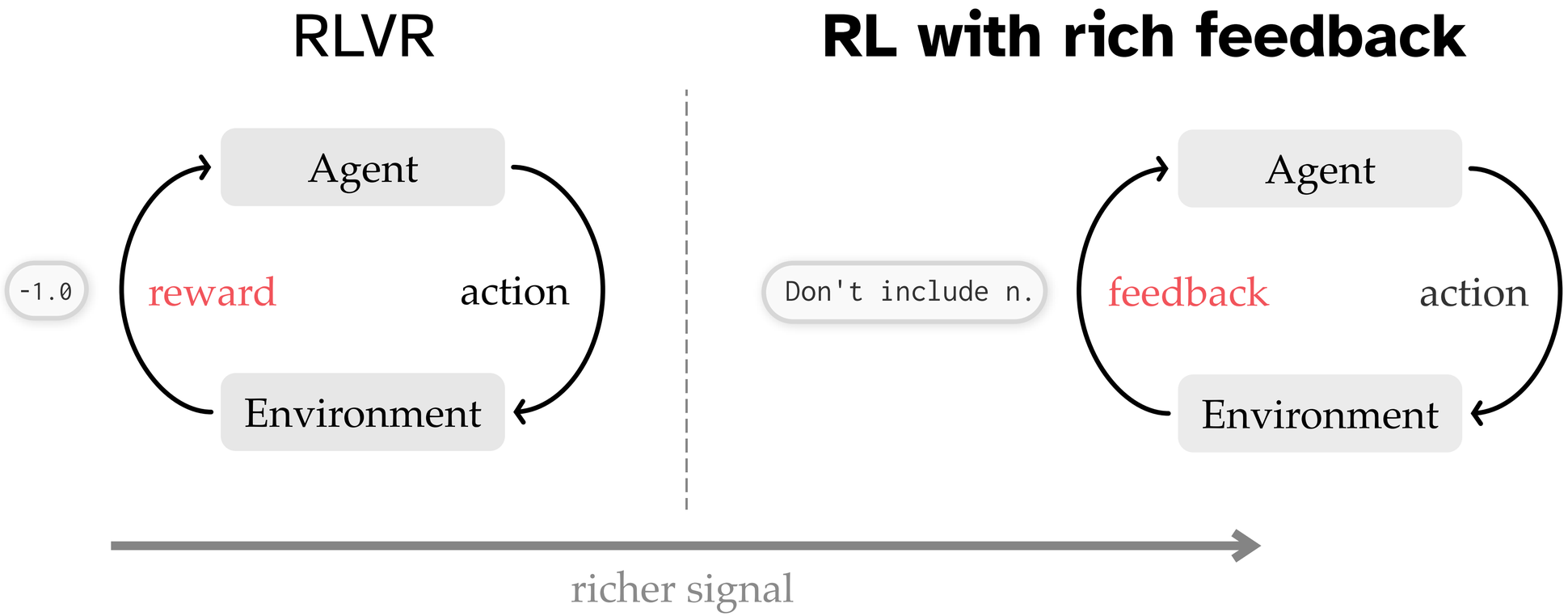
Question:

Write a python function that returns all numbers from 1 to n. Answer briefly.

Answer $y \sim \pi_\theta(\cdot \mid x)$:

```python
def numbers_up_to_n(n):
    return list(range(1, n + 1))
```

incorrect!

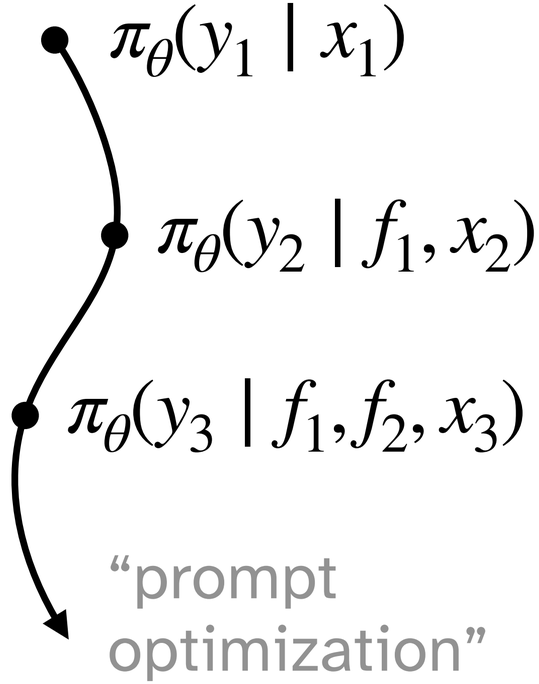


In-context learning

- Enables models to learn from rich feedback, hints, examples, and more.

In-context learning

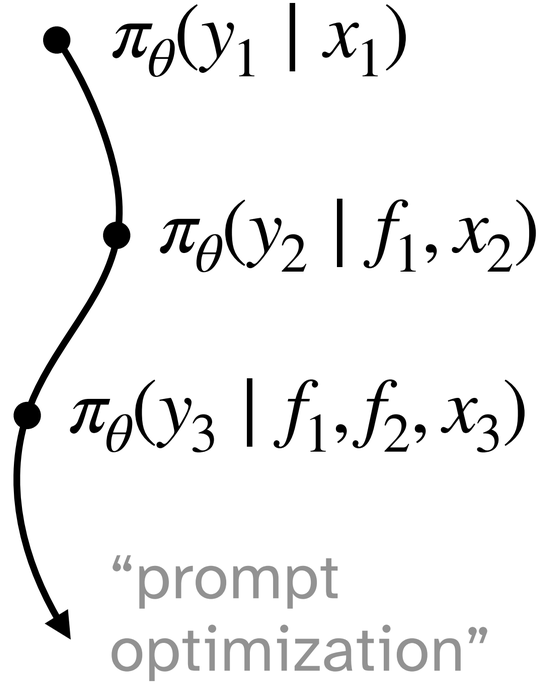
- Enables models to learn from rich feedback, hints, examples, and more.



Prompts x, Responses y, Feedback f

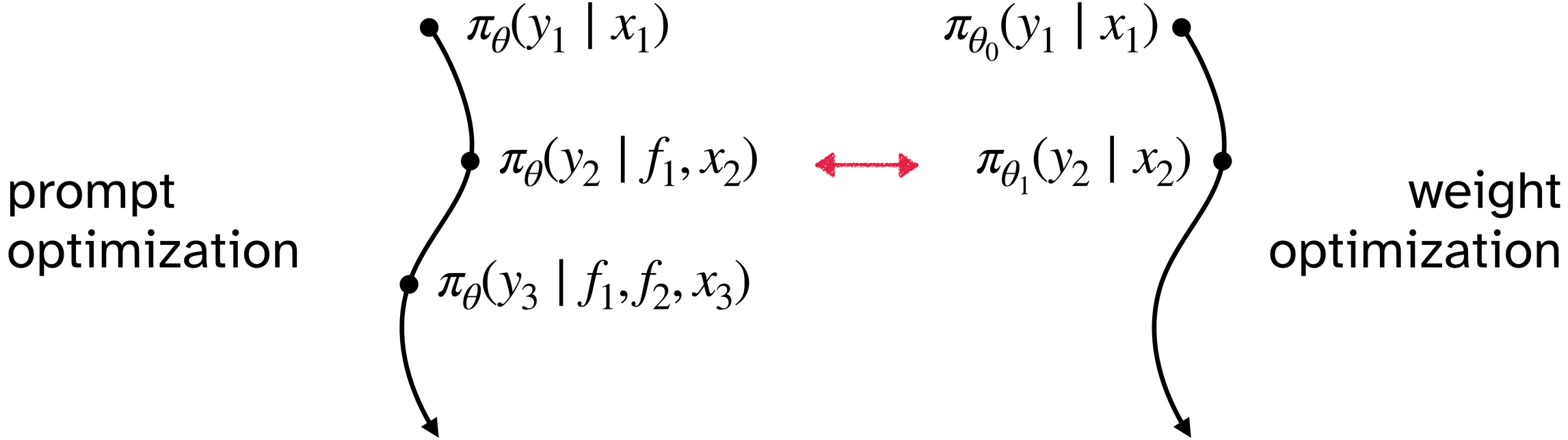
In-context learning

- Enables models to learn from rich feedback, hints, examples, and more.
- **BUT** context grows and learning is transient.



Prompts x, Responses y, Feedback f

In-context learning → Self-distillation



Core idea: Use in-context learning to turn the model into its own teacher, then distill it.

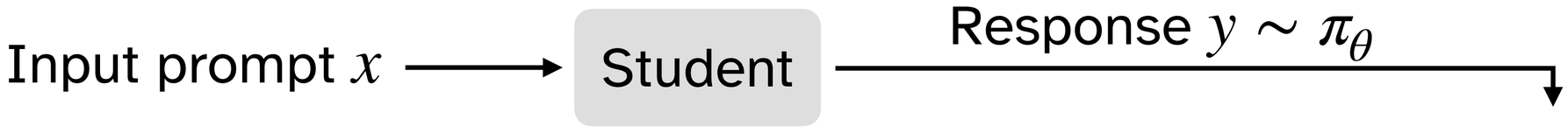
Self-distillation

Prompt x, Response y, Feedback f

Input prompt x

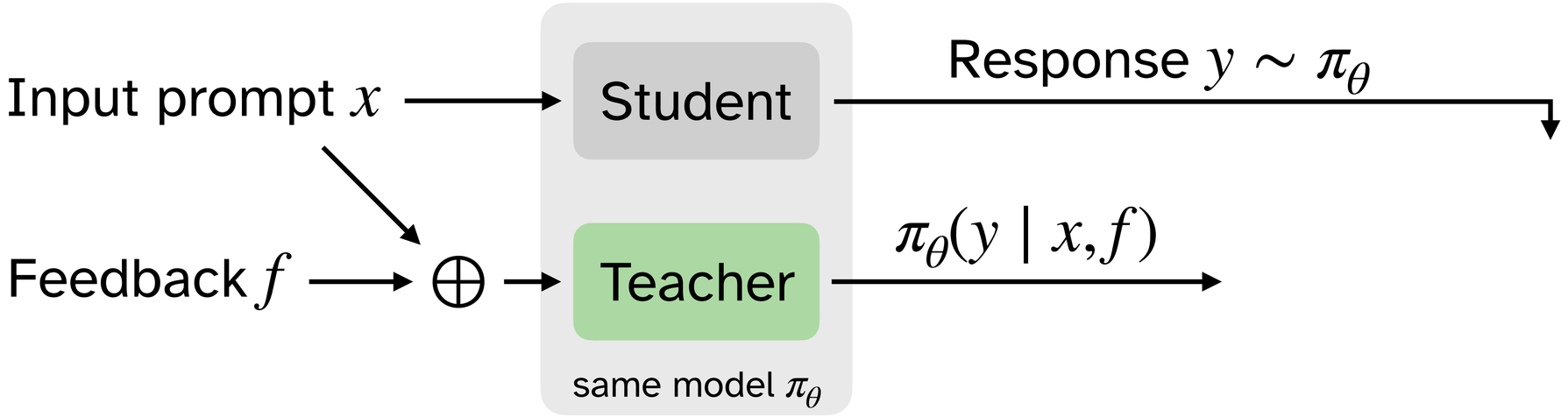
Self-distillation

Prompt x, Response y, Feedback f



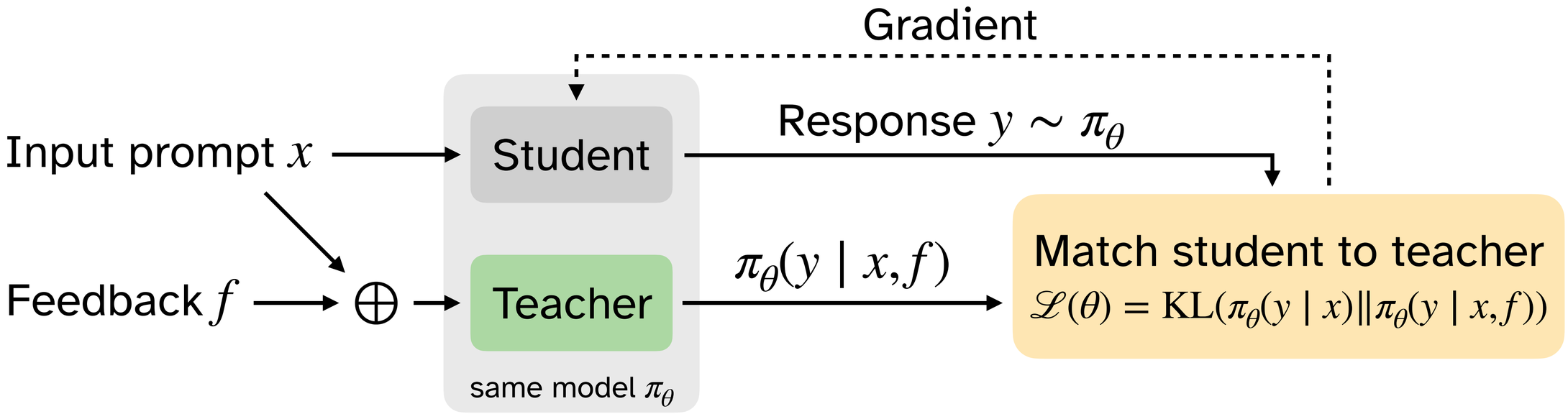
Self-distillation

Prompt x, Response y, Feedback f



Self-distillation

Prompt x, Response y, Feedback f



Two equivalent perspectives

Match student to teacher

$$\mathscr{L}(\theta) = \mathrm{KL}(\pi_\theta(y \mid x) \| \pi_\theta(y \mid x, f))$$

$\Longleftrightarrow$

Token-level advantages

$$A_i(x, y, f) := \log \frac{\pi_\theta(y_i \mid x, f, y_{<i})}{\pi_\theta(y_i \mid x, y_{<i})}$$

On-policy distillation

On-policy RL

Credit assignment in self-distillation

Token-level advantages

$$A_i(x, y, f) := \log \frac{\pi_\theta(y_i \mid x, f, y_{<i})}{\pi_\theta(y_i \mid x, y_{<i})}$$

Question:

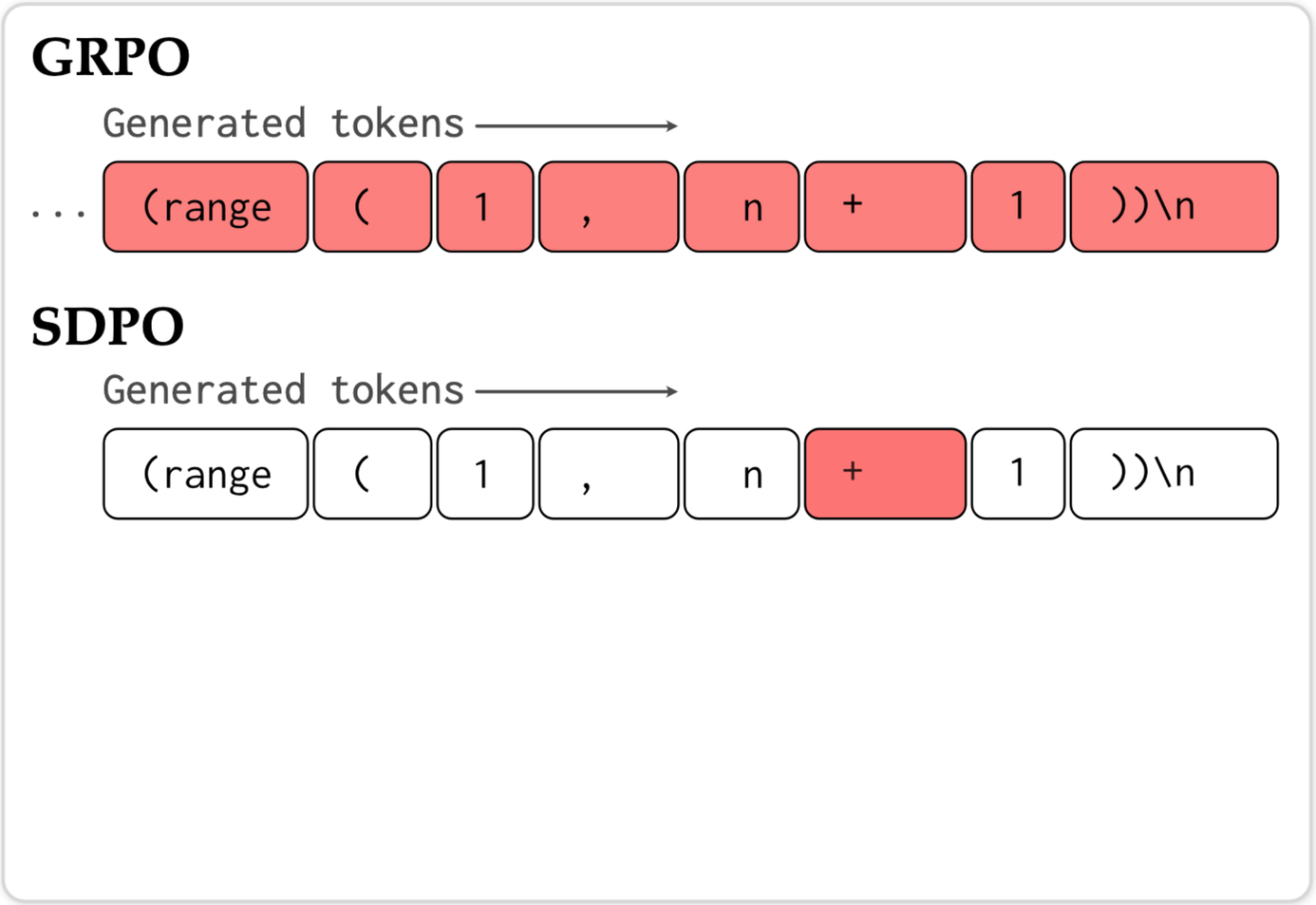
Write a python function that returns all numbers from 1 to n. Answer briefly.

Answer $y \sim \pi_\theta(\cdot \mid x)$:

```python
def numbers_up_to_n(n):
    return list(range(1, n + 1))
```

Feedback:

Don't include n.



Credit assignment in self-distillation

Token-level advantages

$$A_i(x, y, f) := \log \frac{\pi_\theta(y_i \mid x, f, y_{<i})}{\pi_\theta(y_i \mid x, y_{<i})}$$

Question:

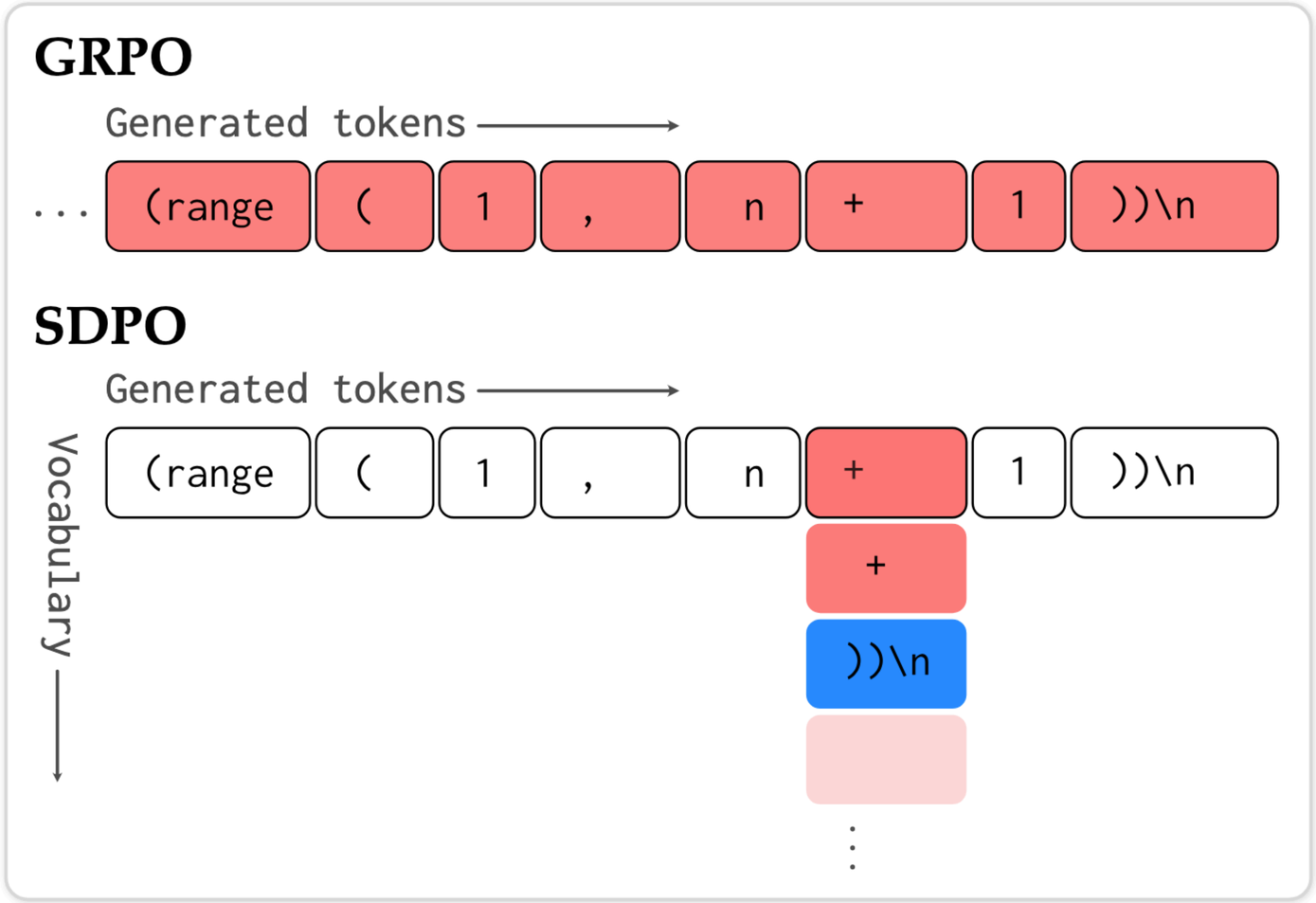
Write a python function that returns all numbers from 1 to n. Answer briefly.

Answer $y \sim \pi_\theta(\cdot \mid x)$:

```python
def numbers_up_to_n(n):
    return list(range(1, n + 1))
```

Feedback:

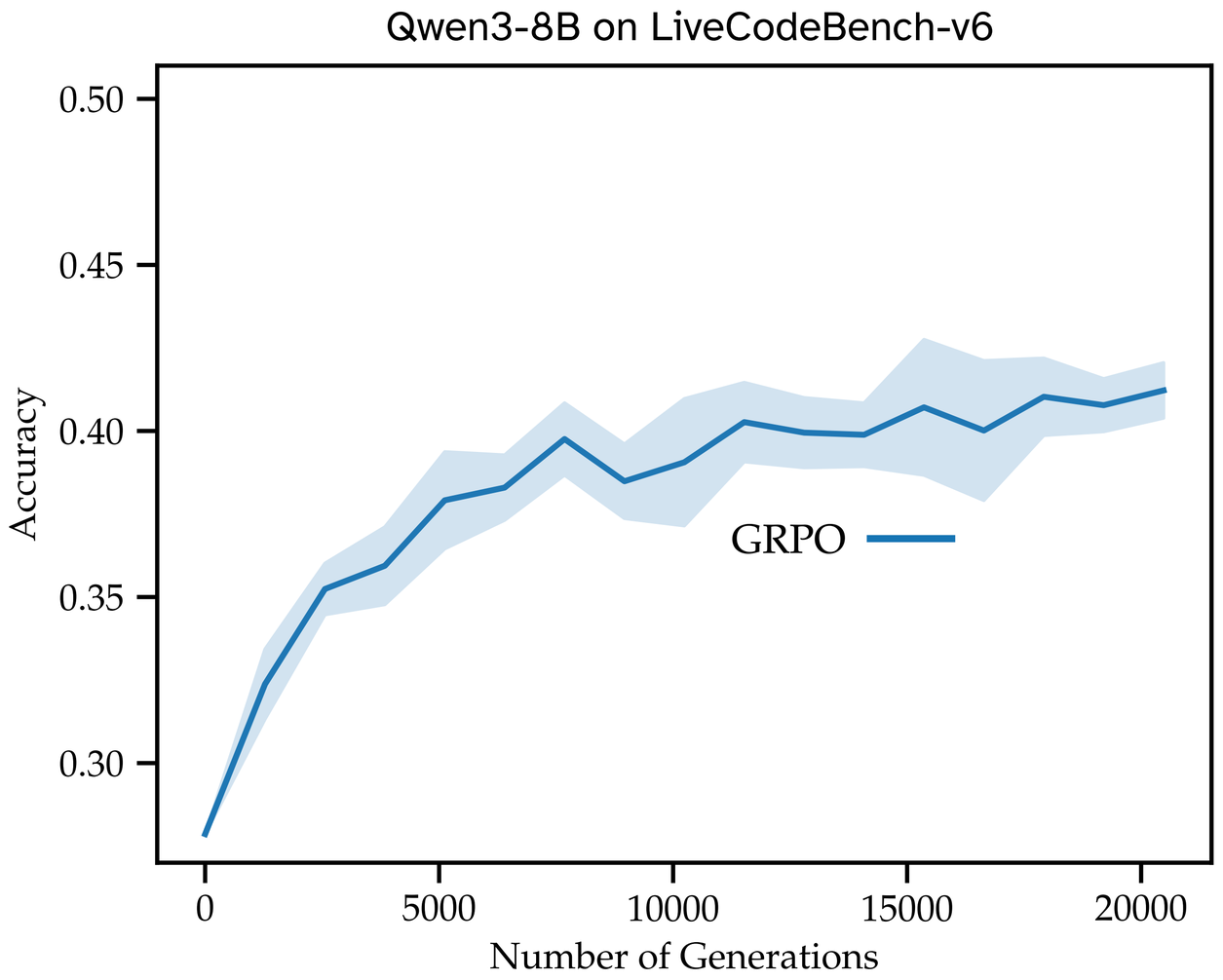
Don't include n.



Results: Learning from execution feedback

Example of feedback

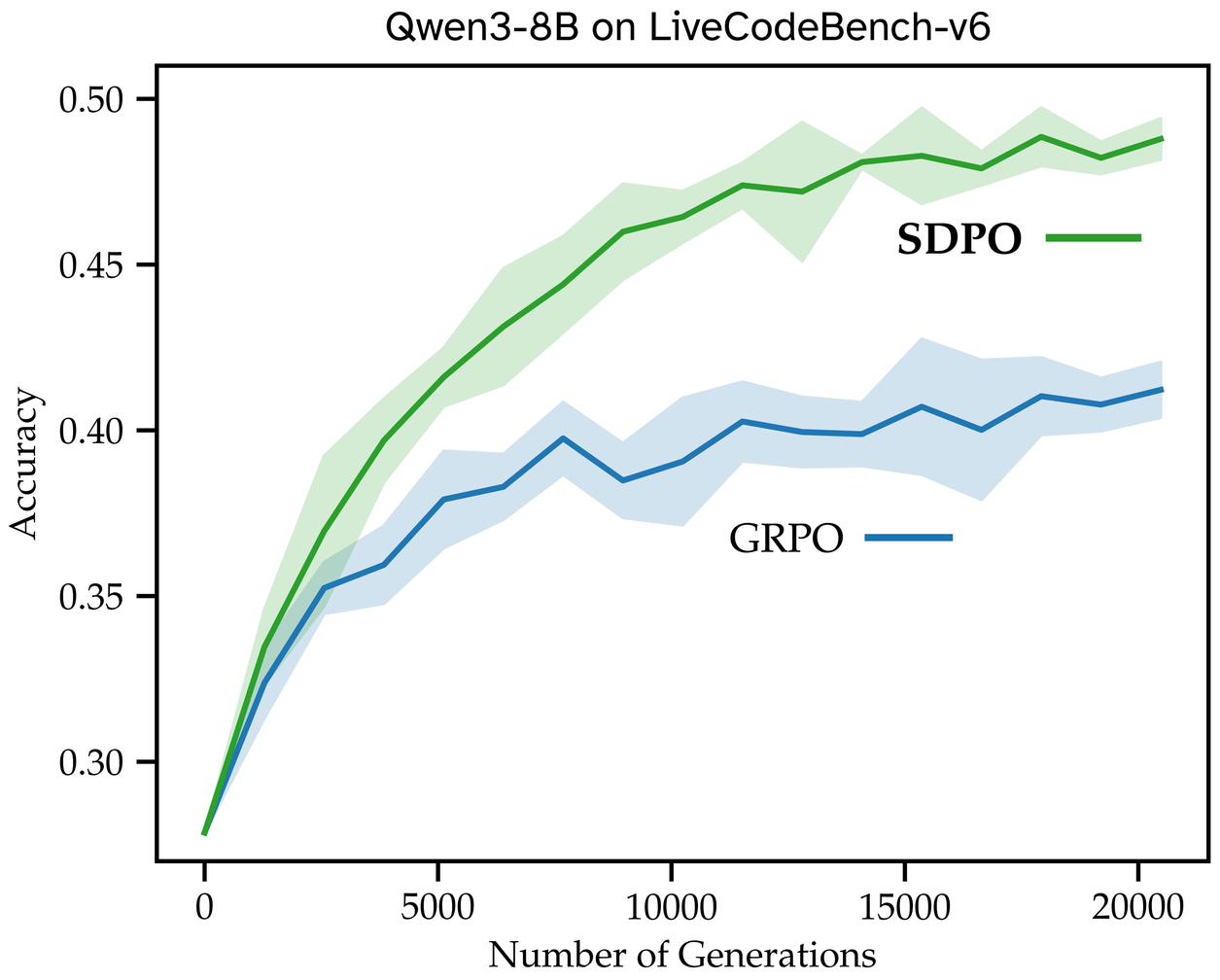
```
Runtime Error
ZeroDivisionError: division by zero
Line 73 in separateSquares (Solution.py)

Last Executed Input
[[26,30,2],[11,23,1]]
```



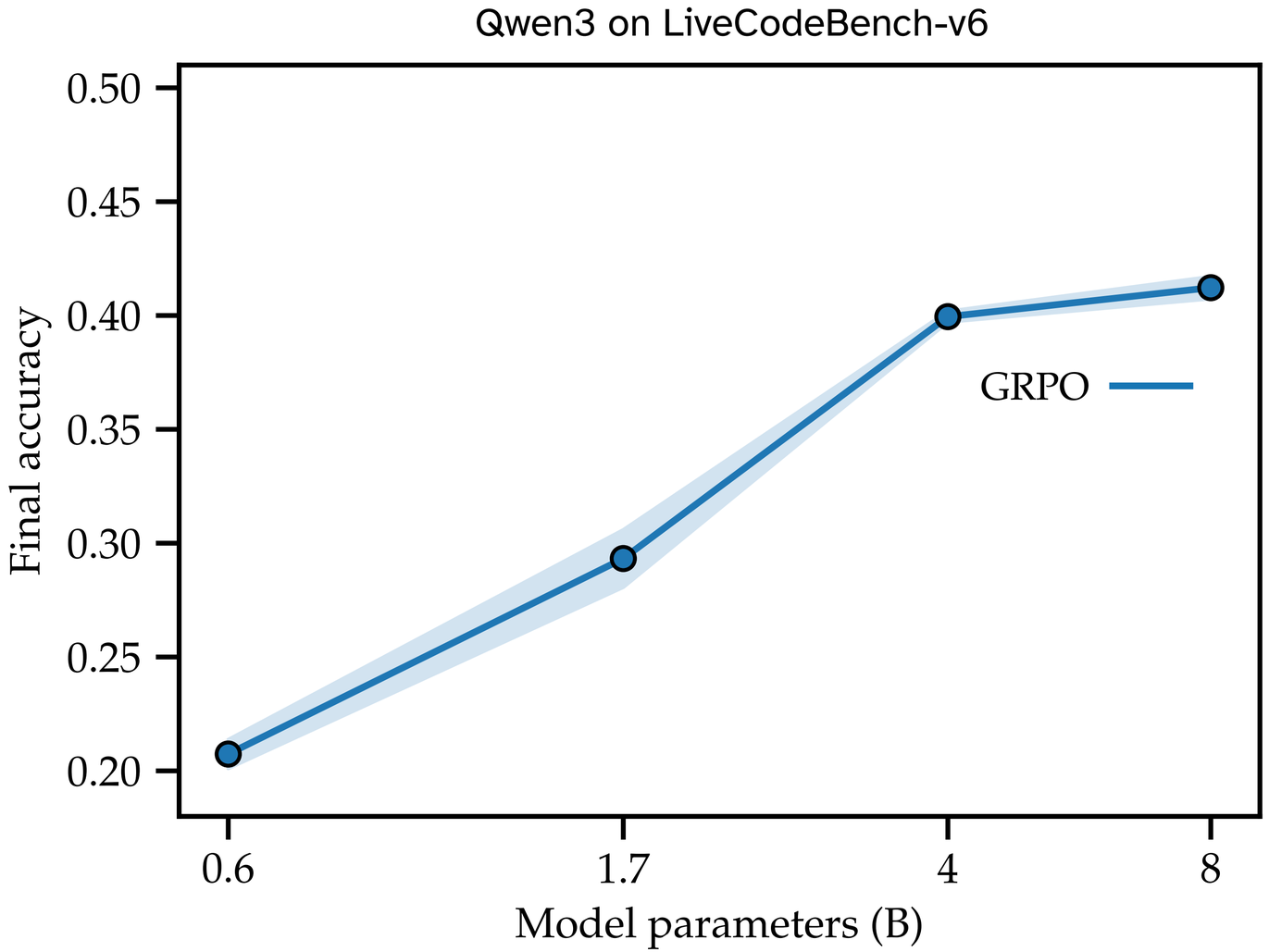
Results: Learning from execution feedback

Example of feedback

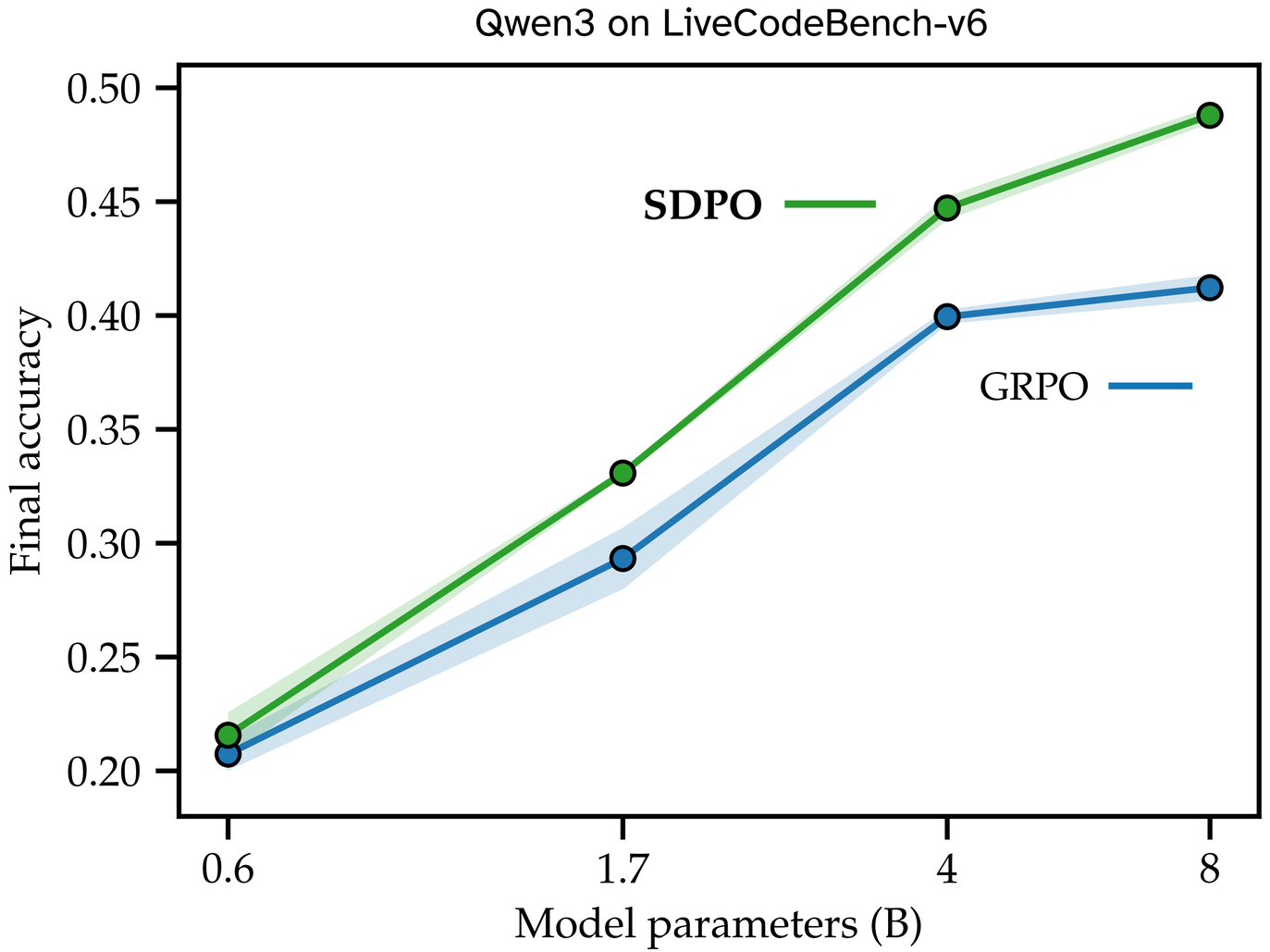
```
Runtime Error
ZeroDivisionError: division by zero
Line 73 in separateSquares (Solution.py)

Last Executed Input
[[26,30,2],[11,23,1]]
```



Self-distillation scales with better models



Self-distillation scales with better models

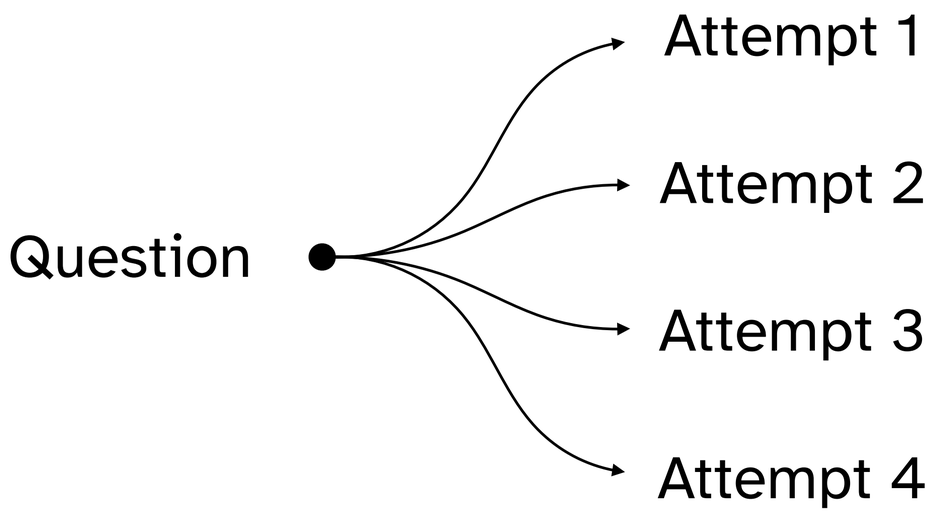


Self-distillation without rich feedback

Can we use self-distillation even with **zero** rich feedback?

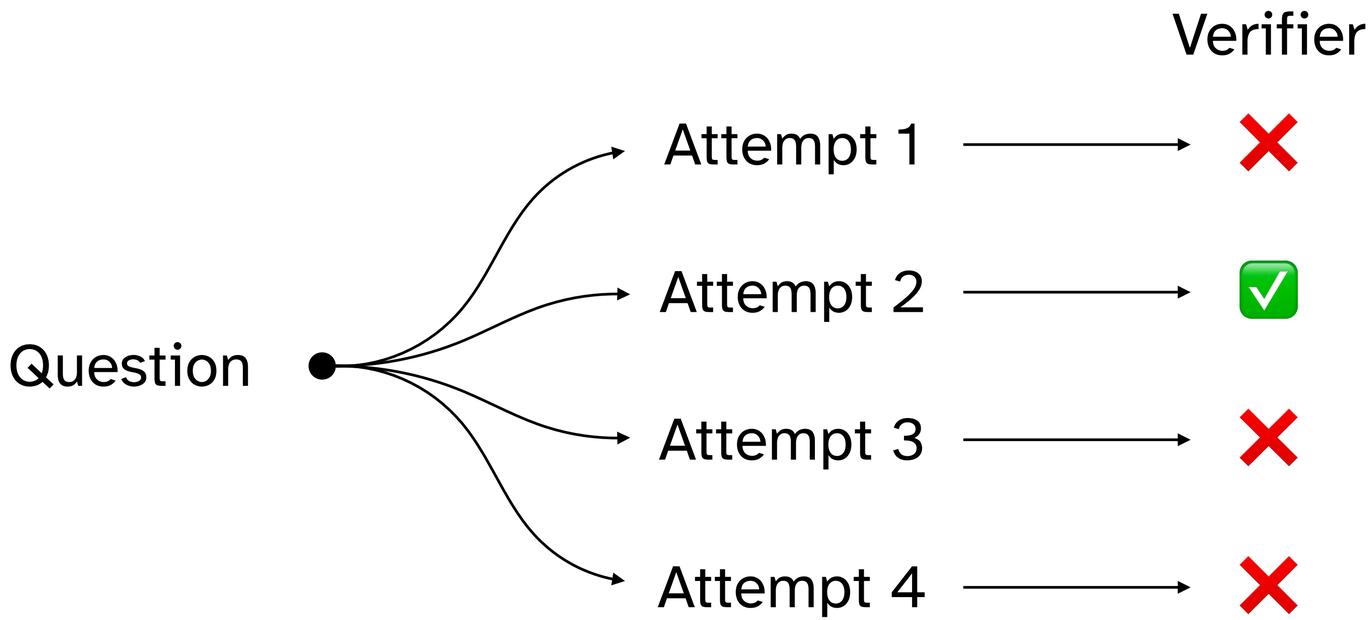
Self-distillation without rich feedback

Can we use self-distillation even with **zero** rich feedback?



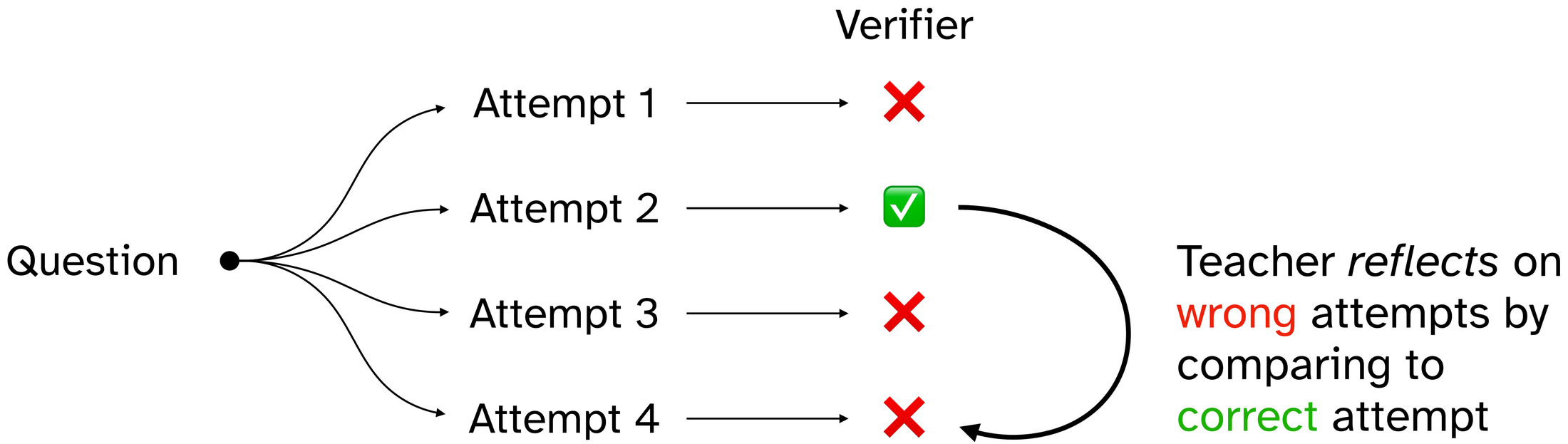
Self-distillation without rich feedback

Can we use self-distillation even with **zero** rich feedback?

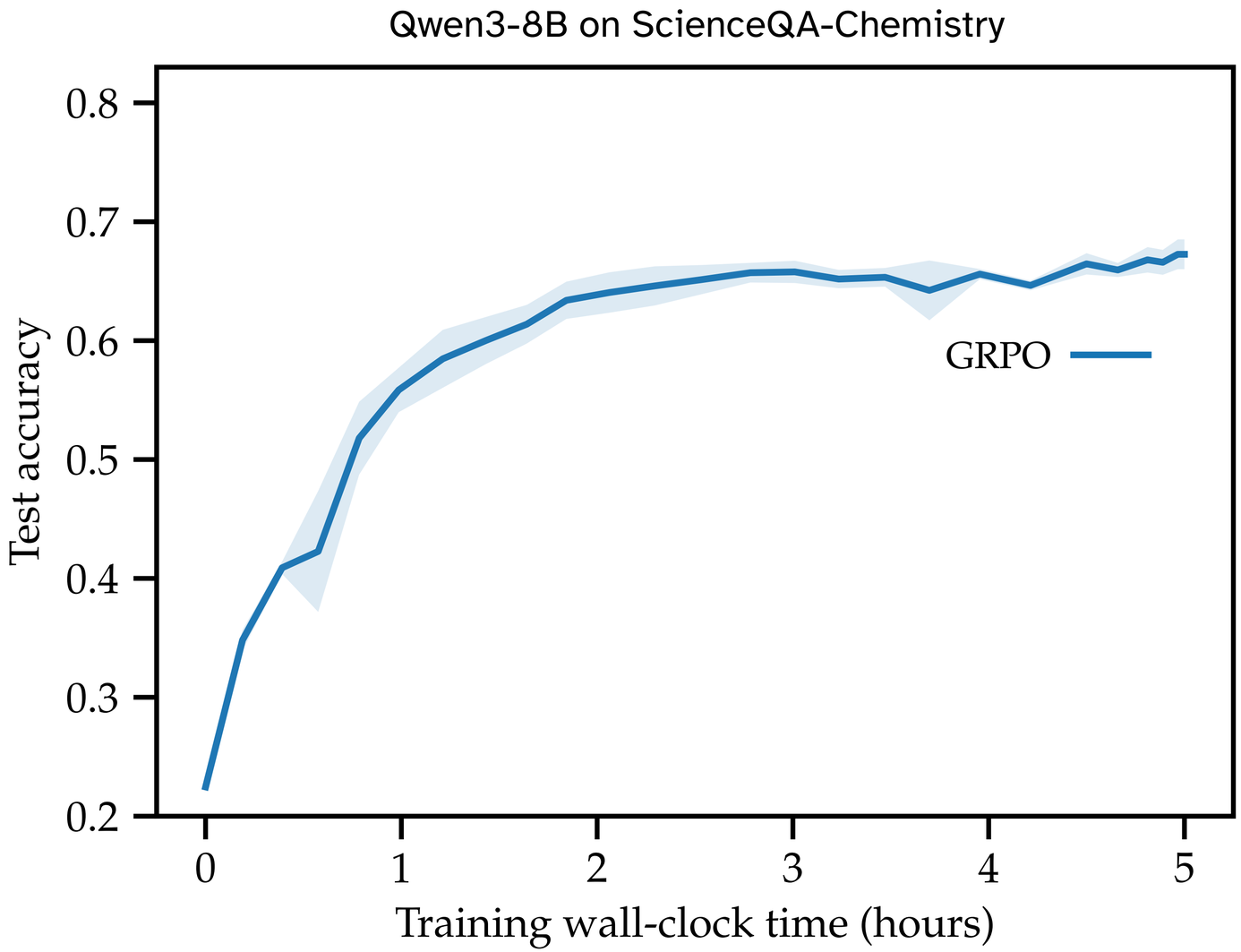


Self-distillation without rich feedback

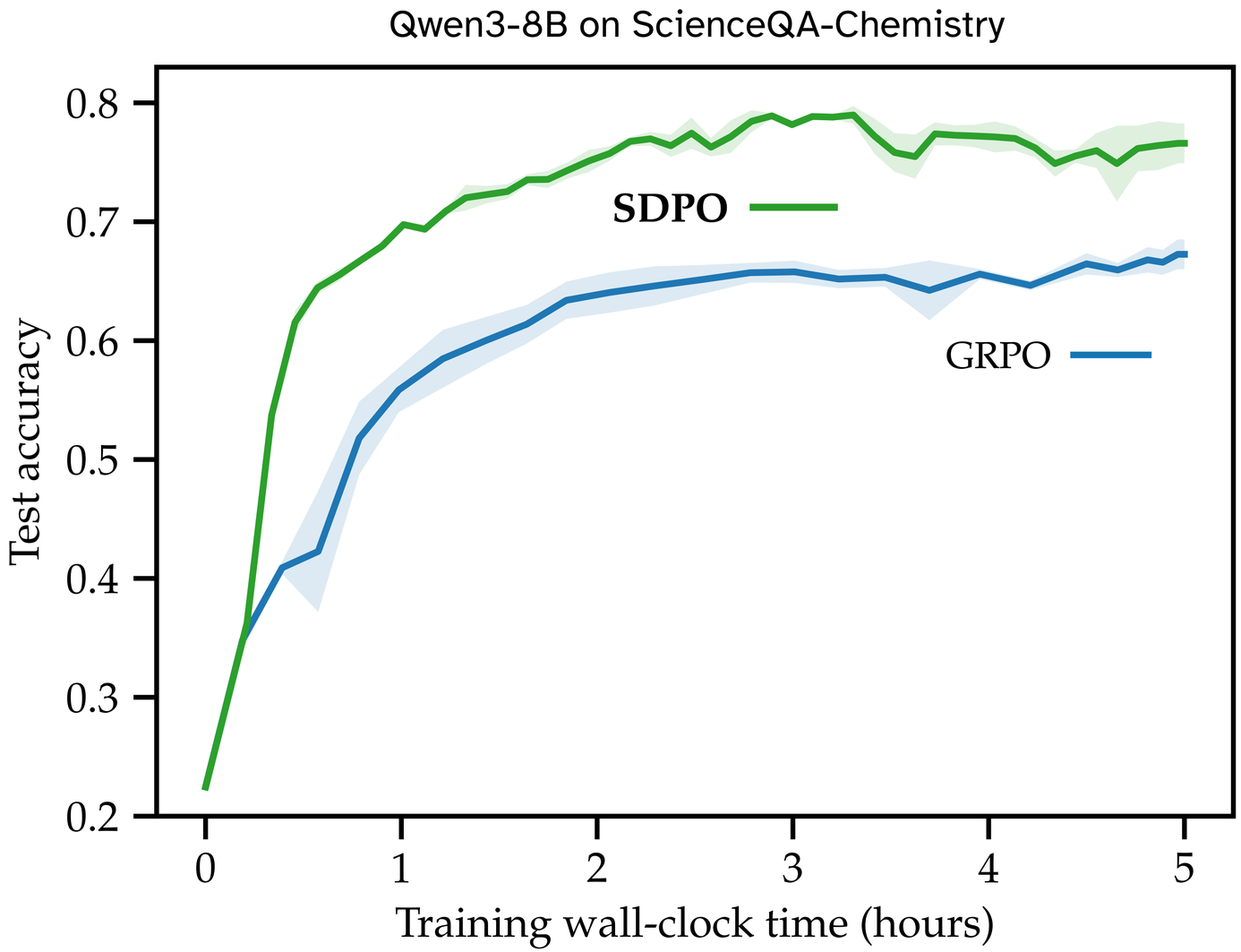
Can we use self-distillation even with **zero** rich feedback?



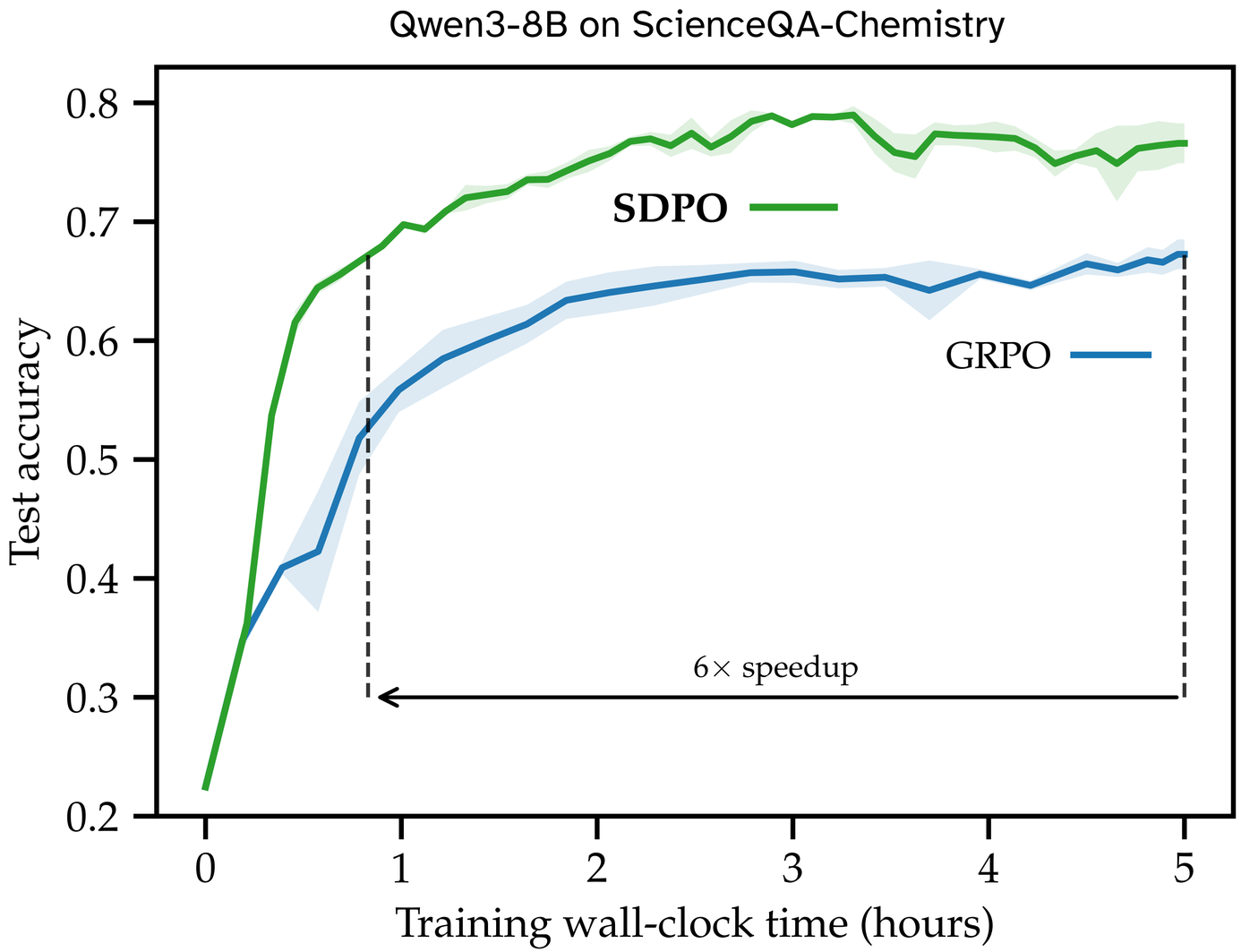
Self-distillation without rich feedback



Self-distillation without rich feedback



Self-distillation without rich feedback



Self-distillation leverages in-context improvement for parametric learning

Check out other applications of self-distillation:

- Shenfeld et al. *Self-Distillation enables Continual Learning*, 2026.
- Buening et al. *Aligning Language Models from User Interactions*, 2026.