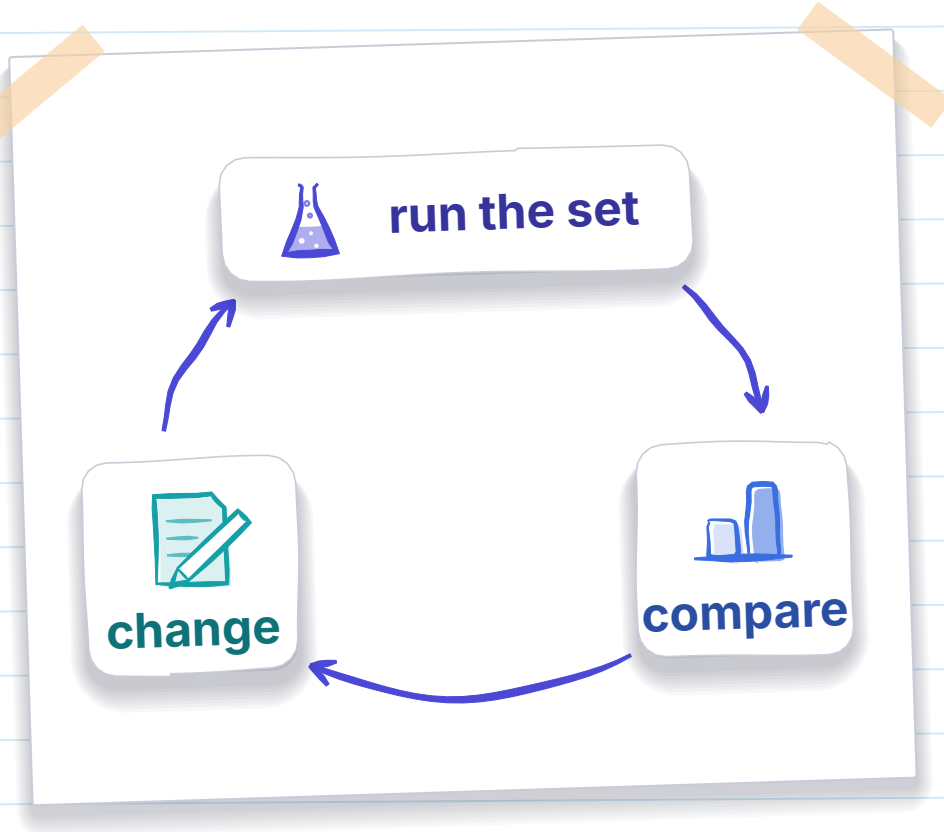
Prof. Aviv Zohar

Semester A, 2026–7

Prompting tricks, and testing them

Chain-of-thought, personas, a benchmark loop



Asking for the working can change the answer

you typed by you

A pen costs $2. A notebook costs $3 more than a pen.
What do two notebooks and a pen cost?

model its reply

$8 

The answer, with no working before it.

Asking for the working can change the answer

you typed by you

A pen costs $2. A notebook costs $3 more than a pen.
What do two notebooks and a pen cost?

model its reply

$8

you added to the same question

Reason step by step.

model its reply

A pen is $2.

A notebook is $2 + $3 = $5.

Two notebooks and a pen: 2 × $5 + $2 = **$12**

This is chain-of-thought prompting: ask for the working.

The steps still need checking.

In 2022, asking for reasoning raised measured scores

Wei et al.: worked examples

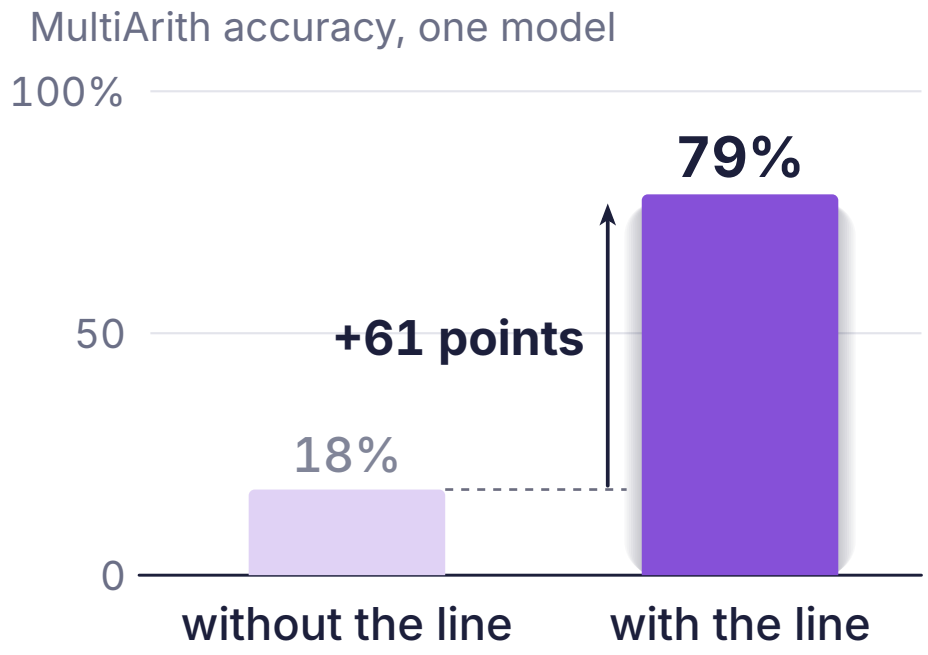
eight of them, each with its reasoning steps

Q: Roger has 5 tennis balls. He buys 2 more cans. Each can has 3 tennis balls. How many does he have now?
A: Roger started with 5 balls. 2 cans of 3 balls is 6 balls. 5 + 6 = 11. The answer is 11.

× 8 examples

Kojima et al.: no examples, one line

`Let's think step by step.`



Same model: a one-line change.

Wei et al., NeurIPS 2022 · Kojima et al., NeurIPS 2022

Reasoning models build chain of thought in

setting **thinking** or **effort**

reasoning model one reply

thinking tokens

A pen is $2, so a notebook is $5. Two notebooks are $10, plus the pen ...

answer

$12

No stock phrase needed to switch it on.

Does “you are an expert” make the answer better?

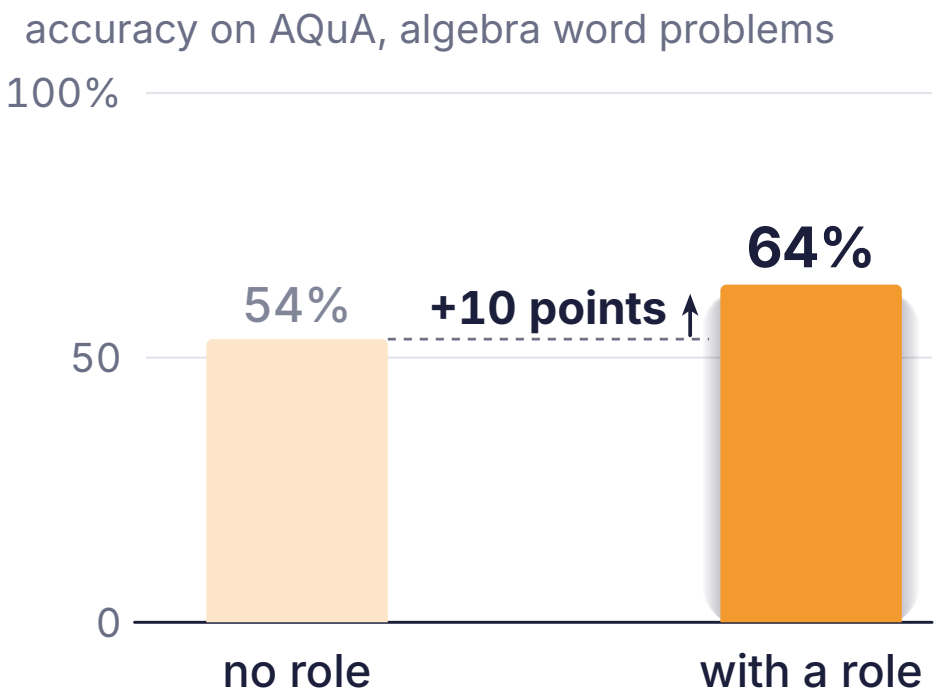


The idea: tell it that it is an expert, and it will answer like one.

Does that hold up when you measure it?

Some studies found a gain from a role

Kong et al.: role-play prompts, ChatGPT



Why might a role help?

a proposed mechanism, not a guarantee

The text it learned to predict holds both:

a careless answer
"just use eval"

a careful, expert one
"validate input first"

may shift the odds toward this text

`You are an expert …`

Later training tunes the model to follow instructions helpfully, so a persona may have less left to add.

Kong et al., NAACL 2024

A useful role spells out the work

you typed by you

```
Act as a senior Python developer.
ⓐ Prefer the standard library.
ⓑ Write typed code.
```

What you can check in the code:

ⓐ only standard-library imports

ⓑ type hints on the function

“senior”: nothing to check

titles.py

```python
from pathlib import Path ⓐ

                         ⓑ
def titles(path: Path) -> list[str]:
    text = path.read_text()
    return text.splitlines()
```

Checkable, whatever the label says.

A role says what kind of work you want

The same code, reviewed by		The review looks at
`You review code for a first-year course.`	→	plain explanations for a beginner
`You are a UX reviewer.`	→	error messages, labels: what users see
`You are a senior developer.`	→	architecture and maintainability
`You are a world-class genius programmer.`	→	a bigger title, not a different job

It may have helped earlier models. Today it probably won't help much.

Prompting is something you test

Worth it for a prompt you reuse: in a skill, or sent by your software many times.

benchmark — agreed answers

1	pen and notebooks	**12**
2	apples in 3 bags	**18**
3	train, 95 minutes	**5:15**
4	6 pizzas, 3 kids	**2 each**
5	two laps, speed	**40**

Decide what counts as correct, and write it down, before the first run.

case 1, in full

prompt

A pen costs $2. A notebook costs $3 more than a pen. What do two notebooks and a pen cost?

agreed answer

12

counts as correct

a final answer of 12, $12 or 12 dollars

Prompting is something you test

benchmark — agreed answers

pen and notebooks	12
apples in 3 bags	18
train, 95 minutes	5:15
6 pizzas, 3 kids	2 each
two laps, speed	40



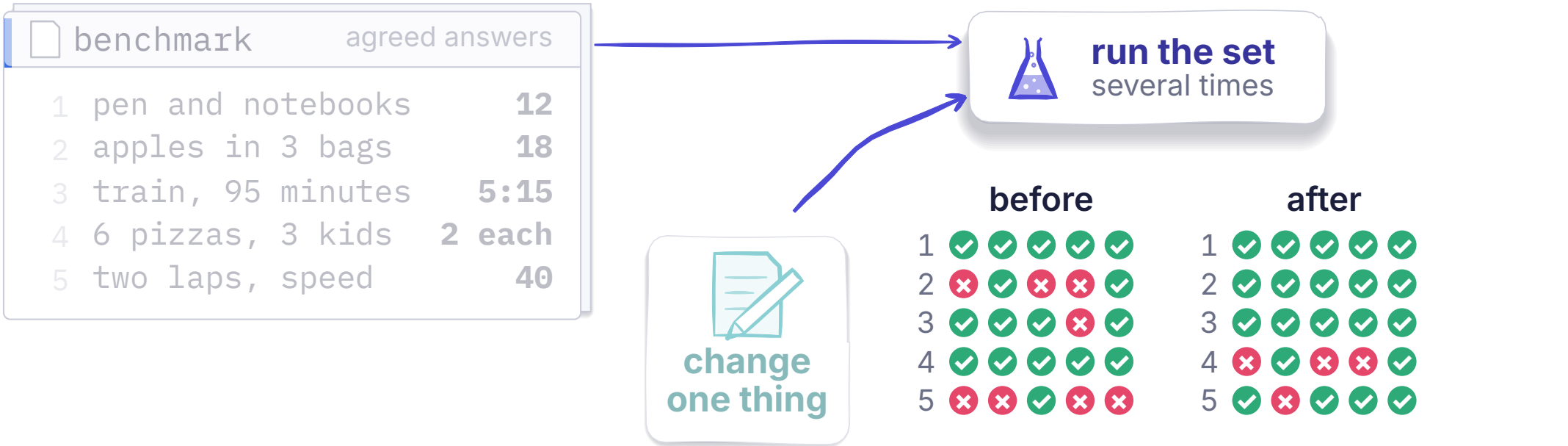
change one thing

prompt.txt — one line added

```
Solve the word problem.
Give the final number.
+ Reason step by step.
```

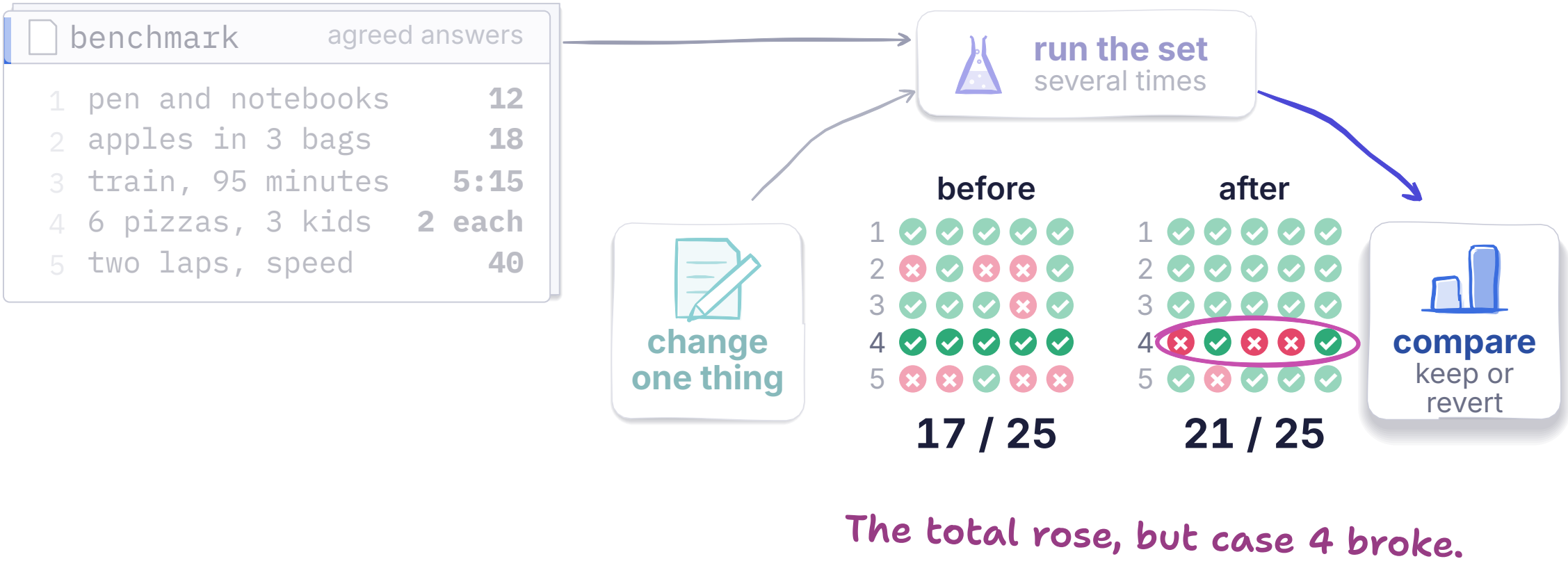
Three changes at once: which one helped?

Prompting is something you test



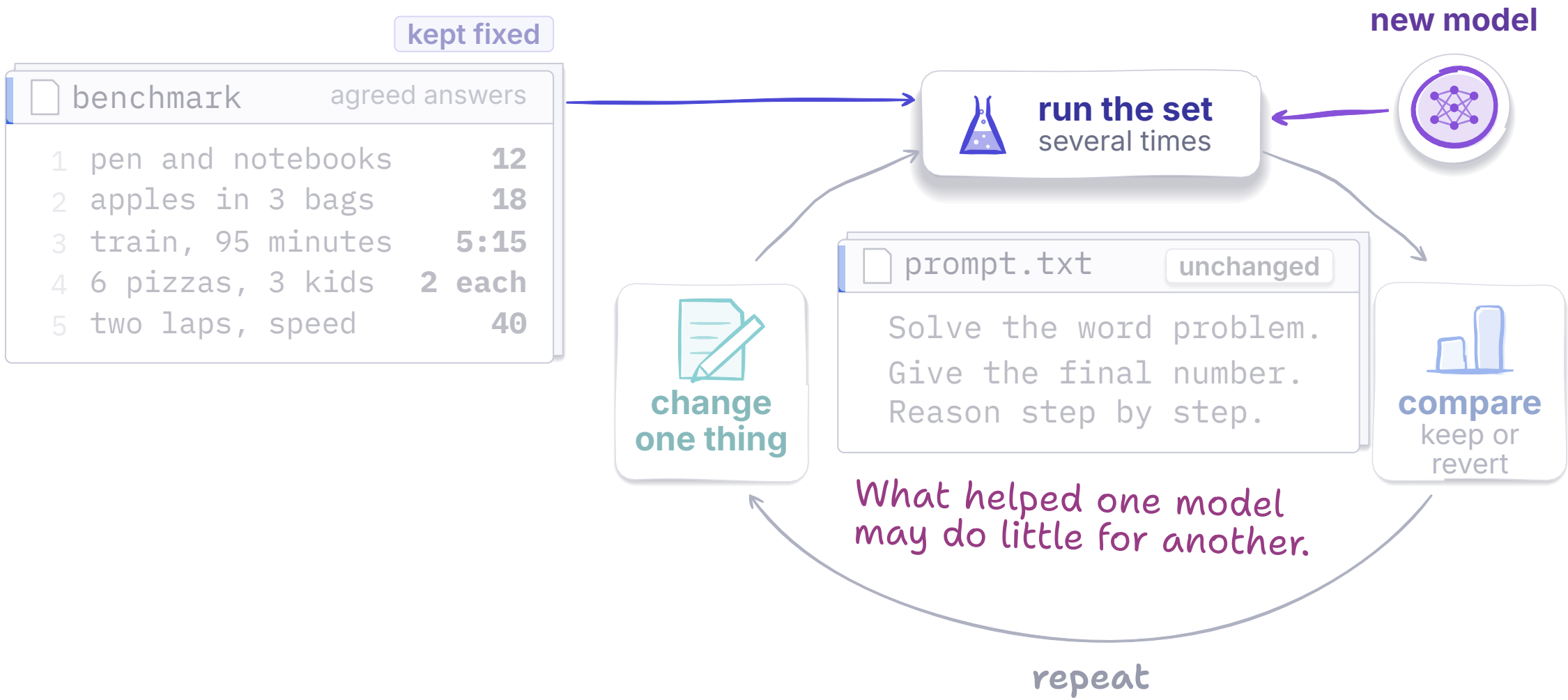
5 cases × 5 runs · same model, same settings

Prompting is something you test



The total rose, but case 4 broke.

Prompting is something you test

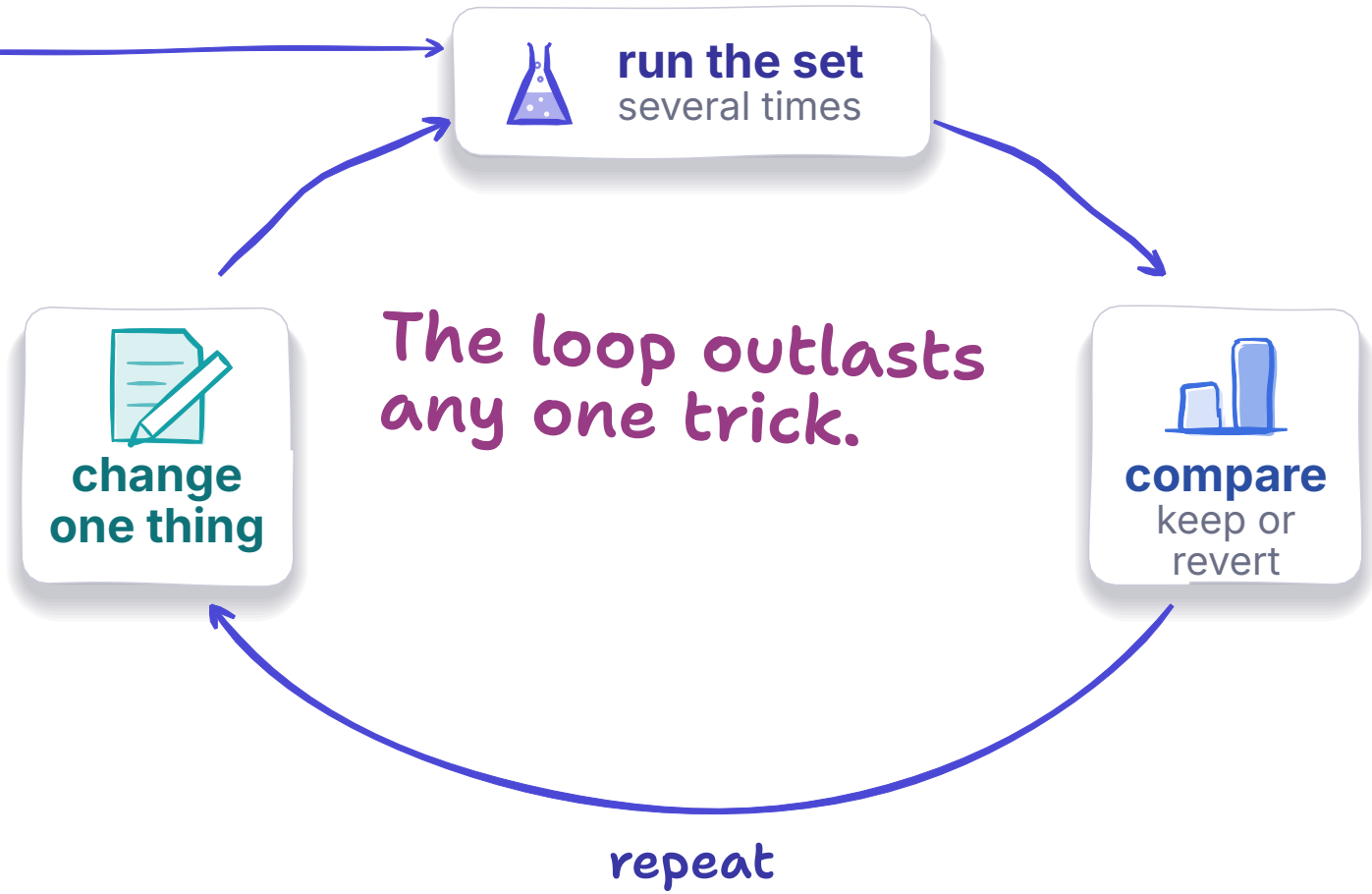


Prompting is something you test

kept fixed

benchmark — agreed answers

1	pen and notebooks	**12**
2	apples in 3 bags	**18**
3	train, 95 minutes	**5:15**
4	6 pizzas, 3 kids	**2 each**
5	two laps, speed	**40**



What to take into the lab



Chain-of-thought

Helped in the 2022 studies. Reasoning models already write their working, so check what an added request for steps changes.

What to take into the lab

Chain-of-thought

Helped in the 2022 studies. Reasoning models already write their working, so check what an added request for steps changes.

Personas

A role can describe the work or the style. Its effect on accuracy depends on the model and task: test the expert label.

What to take into the lab



Chain-of-thought

Helped in the 2022 studies. Reasoning models already write their working, so check what an added request for steps changes.

Personas

A role can describe the work or the style. Its effect on accuracy depends on the model and task: test the expert label.



Testing

Choose a benchmark, change one thing, and compare repeated runs. In the lab, find what helps and where failures remain.

See you in the lab.

Happy prompting!