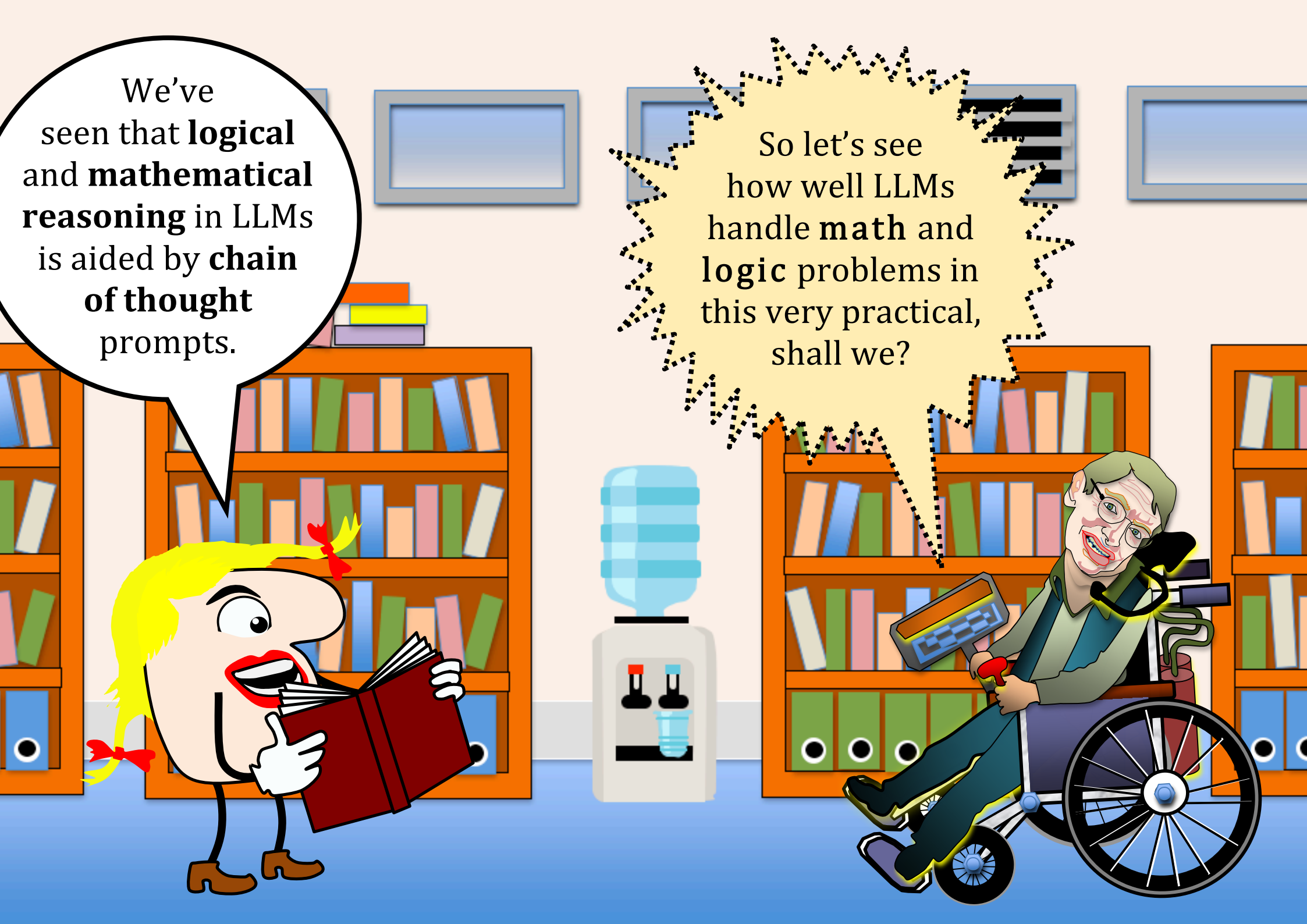
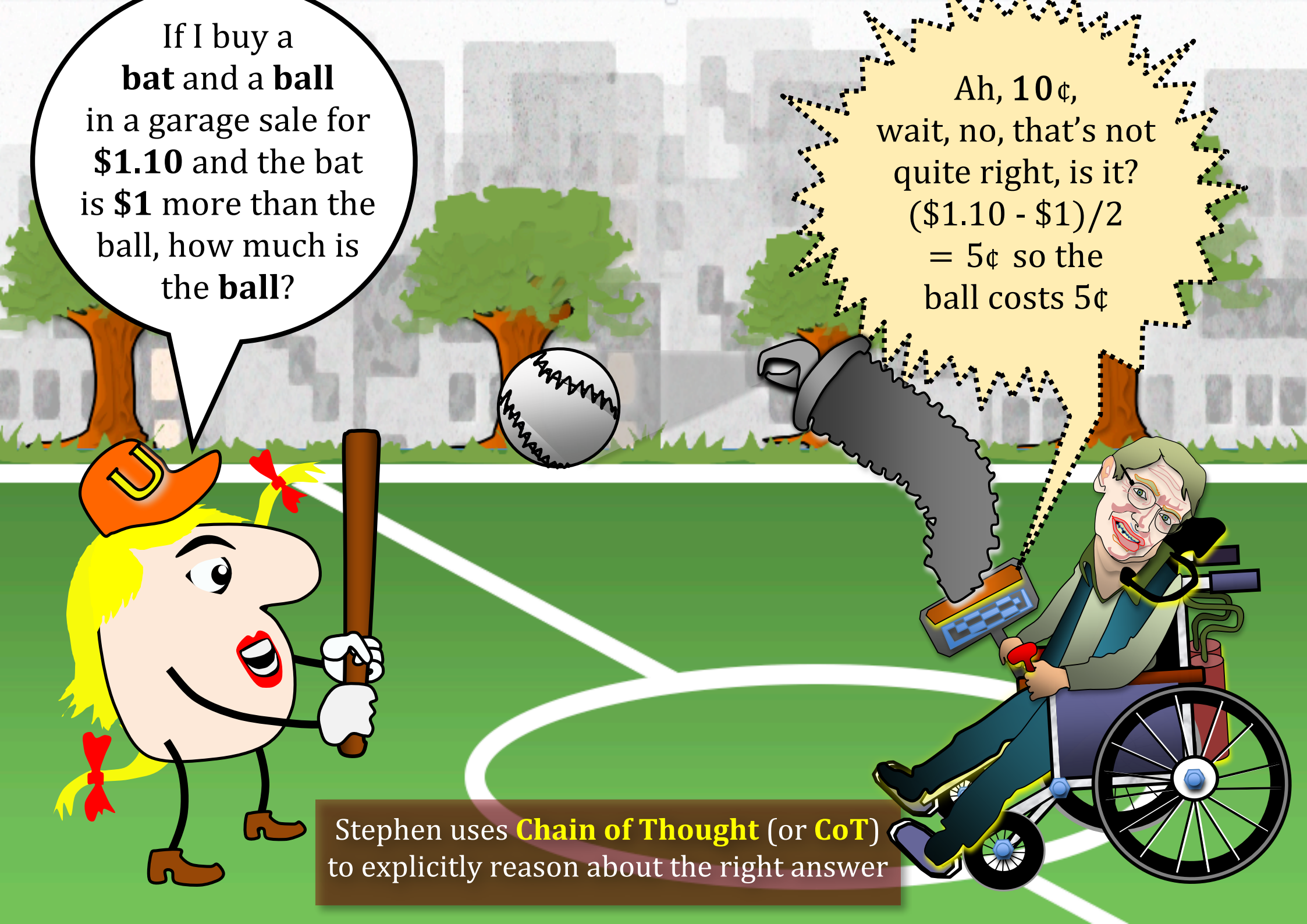
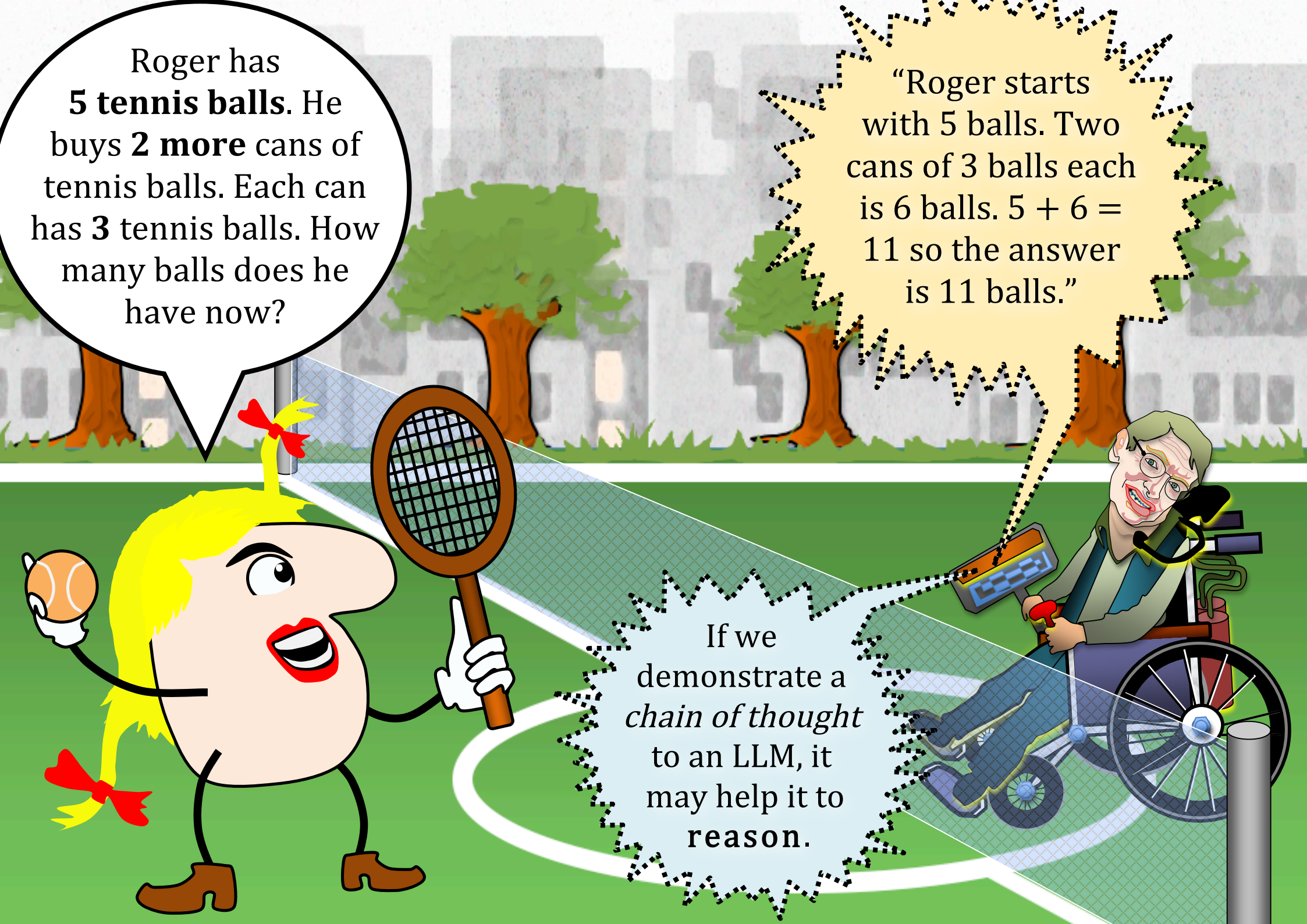
Math Lab: Reasonable Doubt

Tony Veale, 2024


We've seen that **logical** and **mathematical reasoning** in LLMs is aided by **chain of thought** prompts.
So let's see how well LLMs handle **math** and **logic** problems in this very practical, shall we?


If I buy a **bat** and a **ball** in a garage sale for **$1.10** and the bat is **$1** more than the ball, how much is the **ball**?
Ah, **10¢**, wait, no, that's not quite right, is it? ($1.10 - $1)/2 = 5¢ so the ball costs 5¢
Stephen uses **Chain of Thought** (or **CoT**) to explicitly reason about the right answer


Roger has **5 tennis balls**. He buys **2 more** cans of tennis balls. Each can has **3** tennis balls. How many balls does he have now?
"Roger starts with 5 balls. Two cans of 3 balls each is 6 balls. 5 + 6 = 11 so the answer is 11 balls."
If we demonstrate a *chain of thought* to an LLM, it may help it to **reason**.

One-shot with NO Chain of Thought

CONTEXT

Q: *Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many balls does he have now?*

A: *11*

Q: Paddington has 23 oranges. If he uses 20 to make some marmalade and then buys 6 more, how many oranges does he have now?

A: 27 ✗

With **CoT** we **prompt** for a clearly **reasoned** answer.

One-shot with Chain of Thought (CoT)

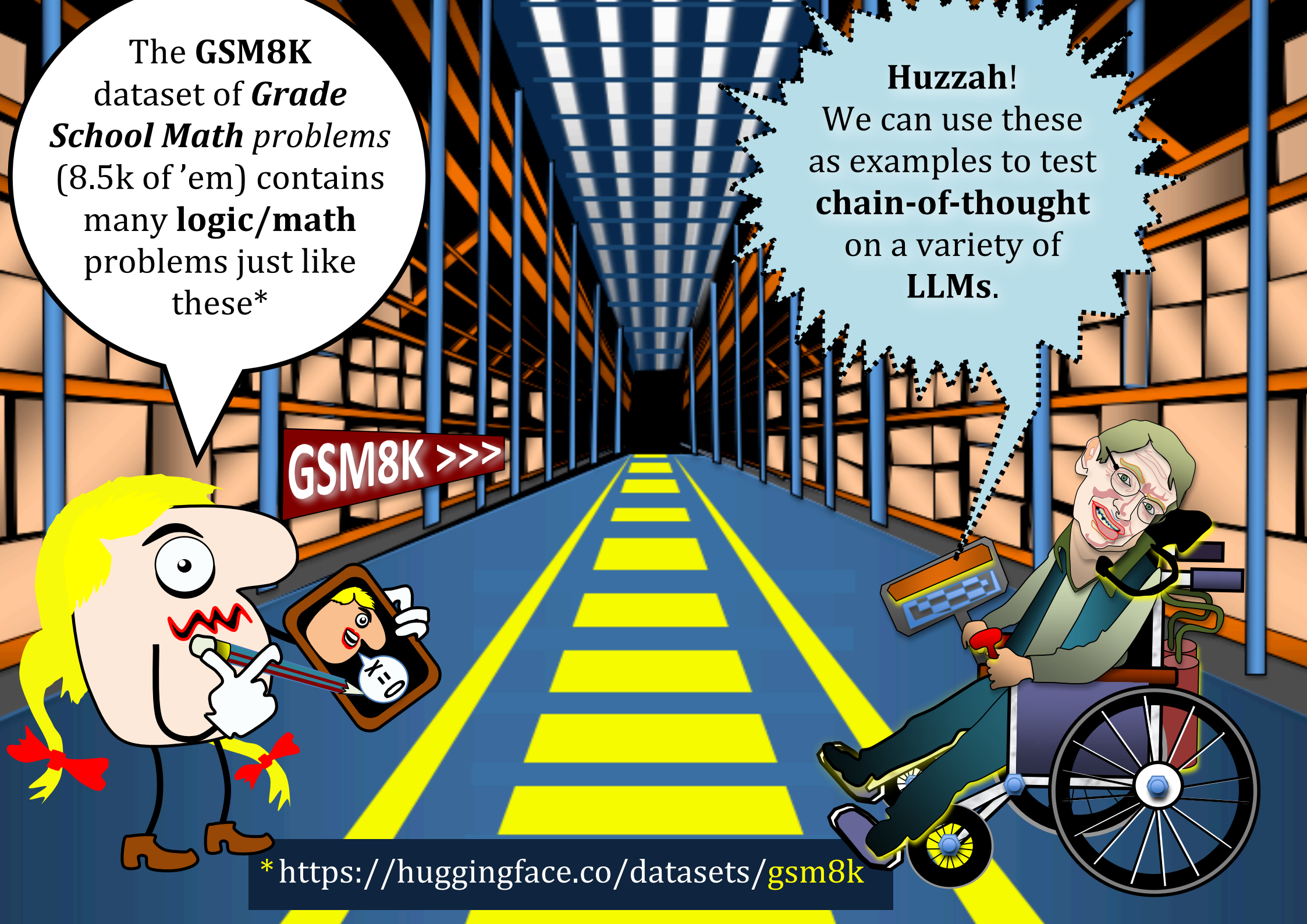
CONTEXT

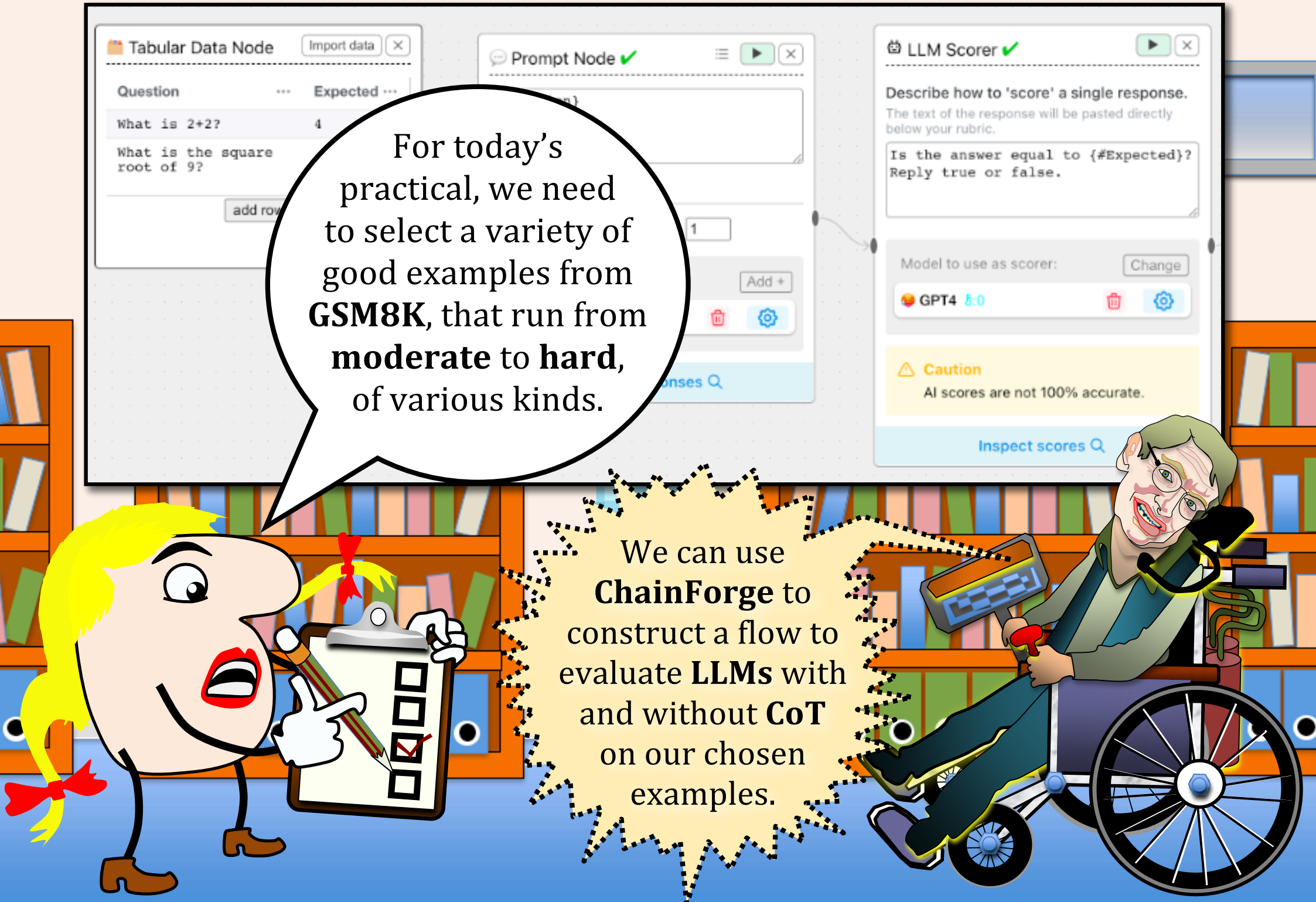
Q: *Roger has 5 tennis balls. He buys 2 more cans of balls. Each can has 3 tennis balls. How many balls does he have now?*

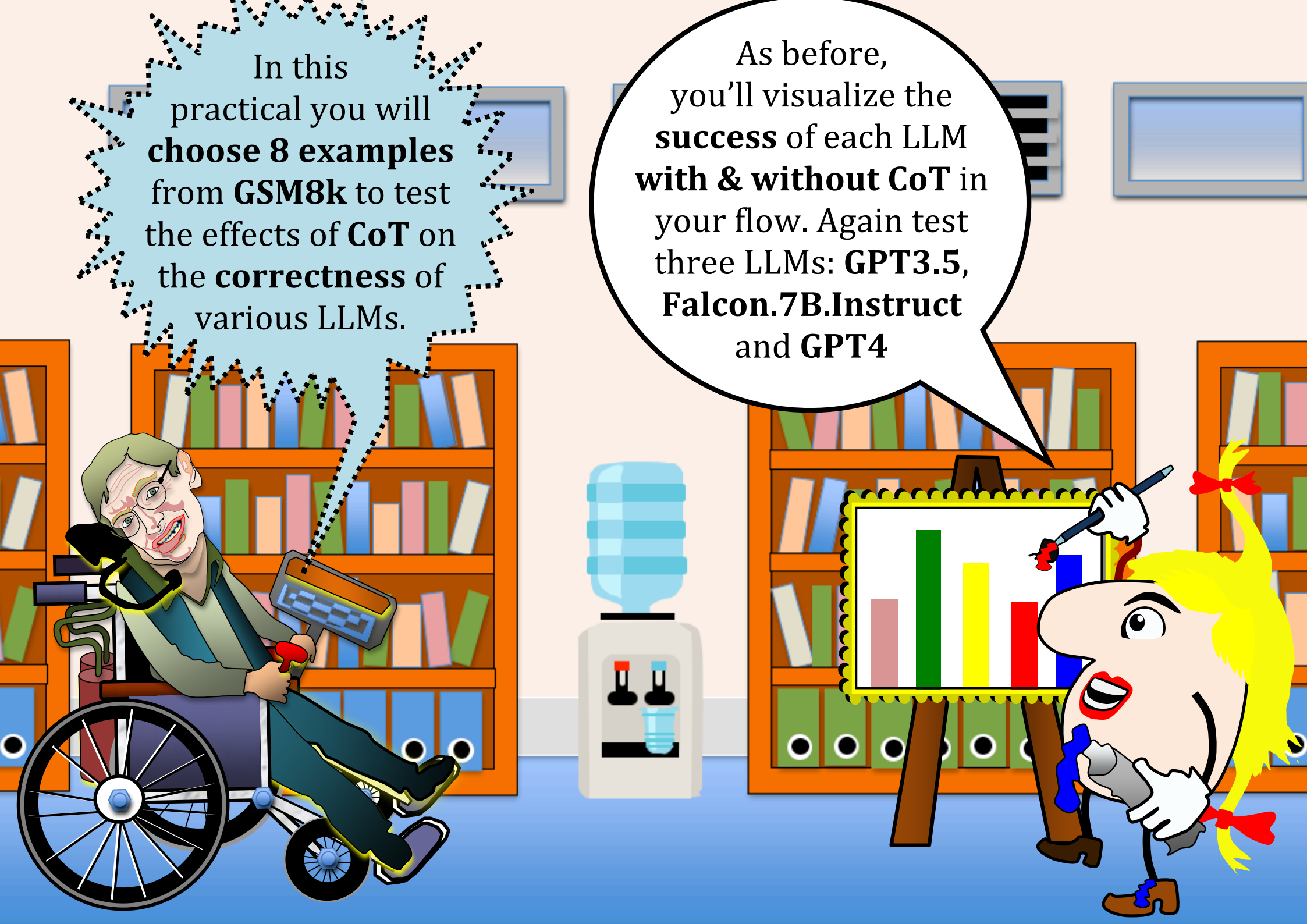
A: *Roger starts with 5 balls. Two cans of 3 balls each is 6 balls. 5 + 6 = 11 so the answer is 11 balls.*

Q: Paddington has 23 oranges. If he uses 20 to make marmalade and then buys 6 more, how many oranges does he have now?

A: Paddington had 23 oranges originally, and used 20 to make marmalade. So he has 23-20 = 3 left. He buys 6 more oranges, so he now has 3 + 6 = 9 oranges. The answer is 9 oranges. ✓

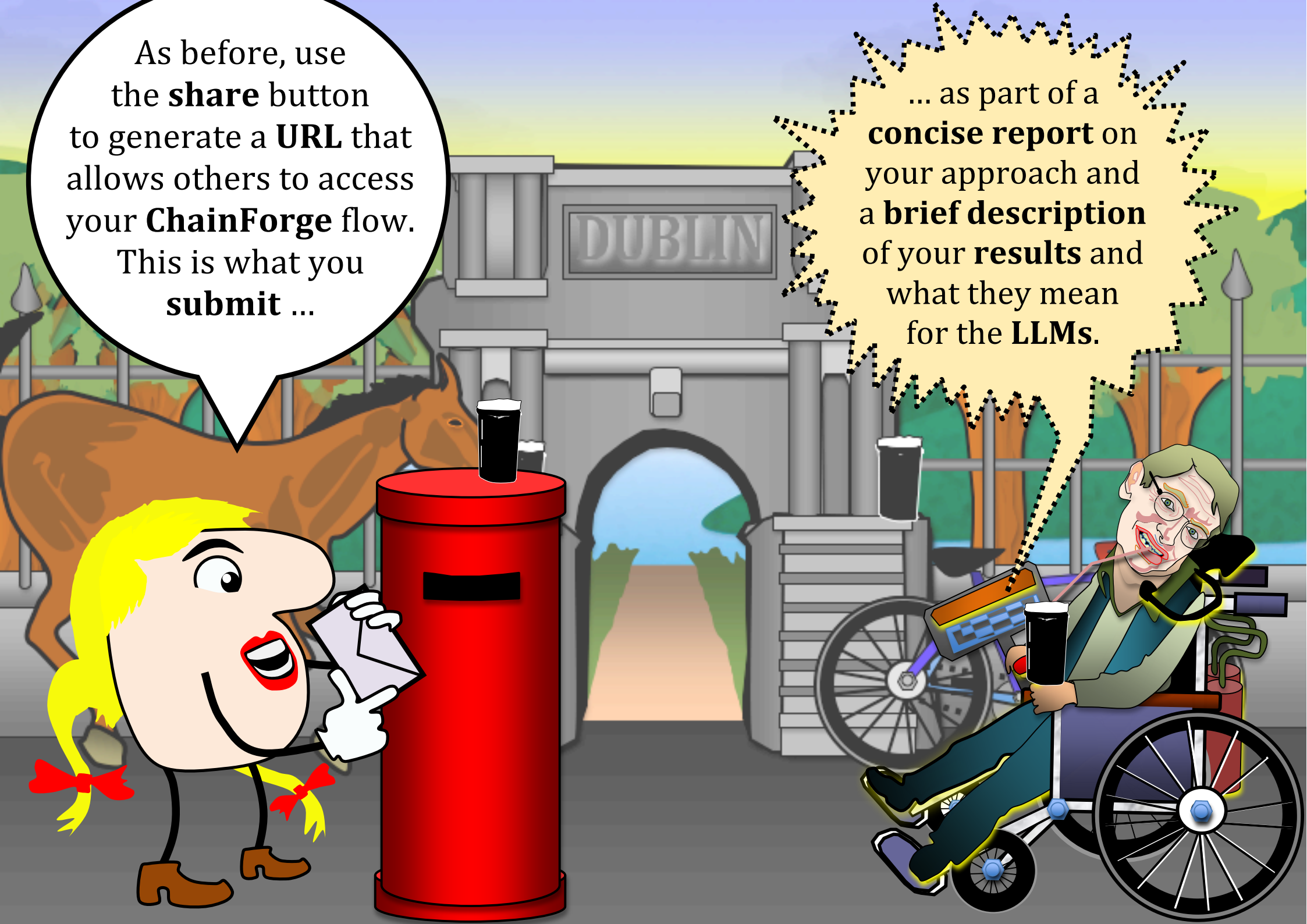

The **GSM8K** dataset of ***Grade School Math*** *problems* (8.5k of 'em) contains many **logic/math** problems just like these*
GSM8K >>>
X=0
Huzzah! We can use these as examples to test **chain-of-thought** on a variety of **LLMs.**
*https://huggingface.co/datasets/gsm8k


Tabular Data Node
Import data
Question
Expected
What is 2+2?
4
What is the square root of 9?
add row
Prompt Node
Add +
LLM Scorer
Describe how to 'score' a single response.
The text of the response will be pasted directly below your rubric.
Is the answer equal to {#Expected}? Reply true or false.
Model to use as scorer:
Change
GPT4
Caution
AI scores are not 100% accurate.
Inspect scores
For today's practical, we need to select a variety of good examples from **GSM8K**, that run from **moderate** to **hard**, of various kinds.
We can use **ChainForge** to construct a flow to evaluate **LLMs** with and without **CoT** on our chosen examples.


In this practical you will **choose 8 examples** from **GSM8k** to test the effects of **CoT** on the **correctness** of various LLMs.
As before, you'll visualize the **success** of each LLM **with & without CoT** in your flow. Again test three LLMs: **GPT3.5**, **Falcon.7B.Instruct** and **GPT4**


Again, don't break the bank! Use **Falcon only** at first for development (it's **free**) before bringing in **GPT3.5** to fine-tune, then **GPT4**.
Quite so. To test **GPT3.5** for free, you can also run your prompts through **ChatGPT** to debug them outside **ChainForge**.


Your efforts will be judged on the **quality** of your flow, **choice** of examples, and **thoroughness** of your evaluation of the LLMs.
Do try to **challenge** the LLMs with good examples to **show** the **benefits** of **CoT**. We aim to **validate CoT** as a **strategy**.


As before, use the **share** button to generate a **URL** that allows others to access your **ChainForge** flow. This is what you **submit** ...
... as part of a **concise report** on your approach and a **brief description** of your **results** and what they mean for the **LLMs**.
DUBLIN