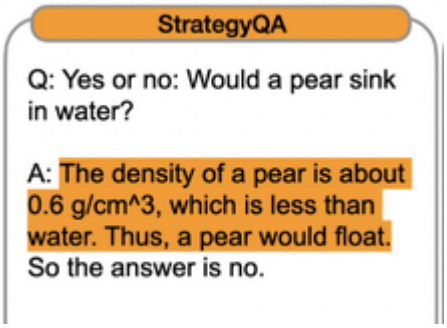
Twitter Thread by David Chapman

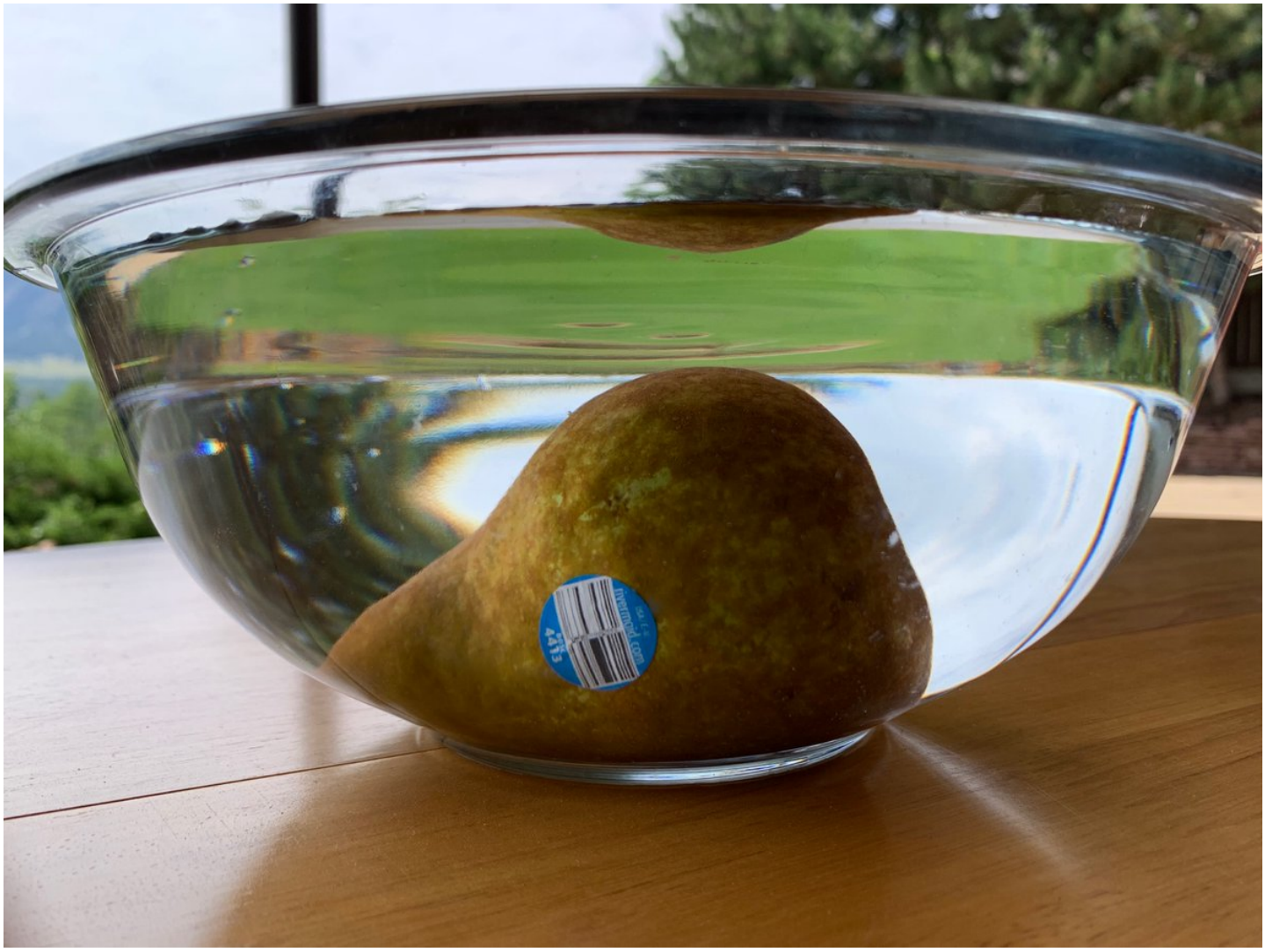
David Chapman
@Meaningness

■ This example from the Google AI blog has been bothering me for ages. It can't possibly be right. The density of living tissue is always nearly 1, and pears feel heavier in the hand than most fruit.



So today I resolved to ¡SCIENCE IT! and bought a pear and can report that this one, at least, sinks in water. (Just barely. Its density must be a hair over 1.)

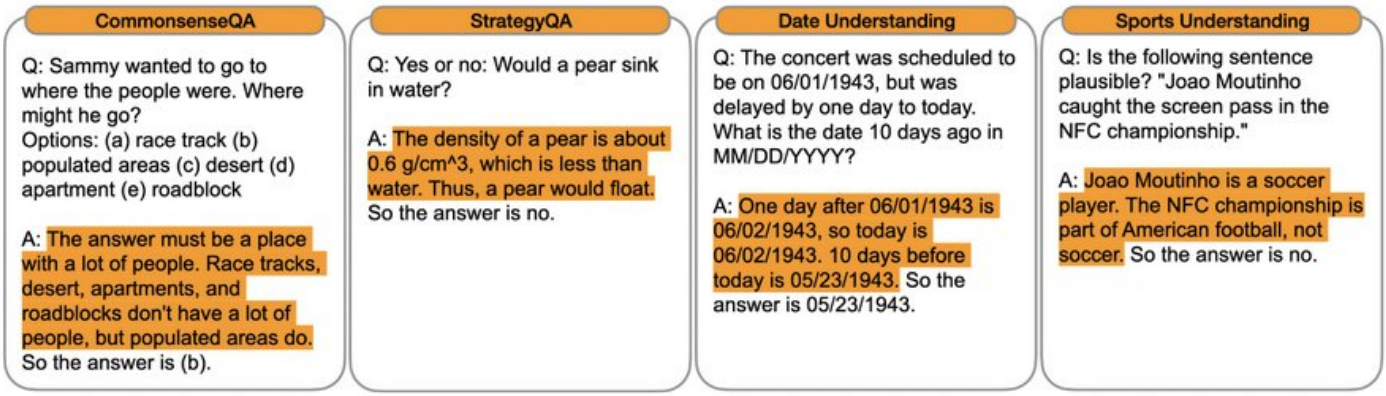
Camera angle isn't great, but best I could do—it's looking through the side of a glass bowl, beneath the surface level.



So what’s going on here? I took the Google AI blog as saying that the answer was generated by one of their AI systems, and that’s been the assumption of other analyses on the internet, but they don’t actually say that! It’s just an example problem. https://t.co/RCdv5KF6NH

Commonsense Reasoning

In addition to arithmetic reasoning, we consider whether the language-based nature of chain of thought prompting also makes it applicable to commonsense reasoning, which involves reasoning about physical and human interactions under the presumption of general background knowledge. For these evaluations, we use the CommonsenseQA and StrategyQA benchmarks, as well as two domain-specific tasks from BIG-Bench collaboration regarding date understanding and sports understanding. Example questions are below:



The problem comes from the StrategyQA benchmark, which I downloaded. Here it is!

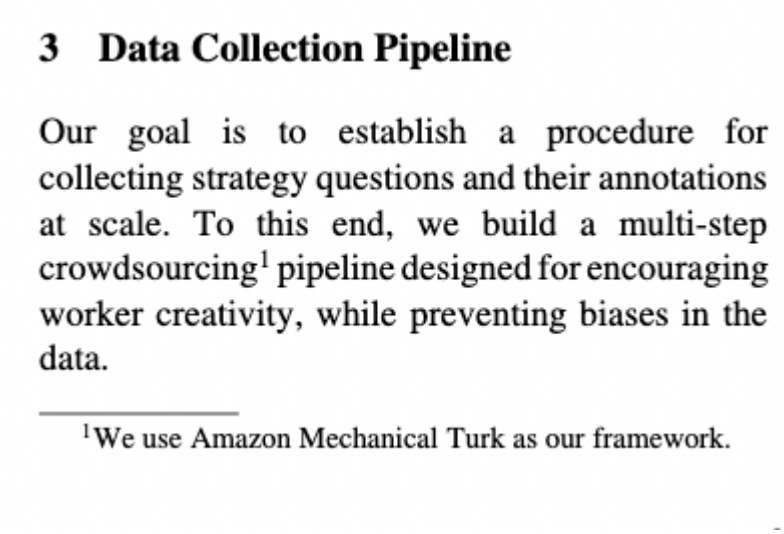
The answer according to the benchmark includes a density claim of 0.59.

The plot thickens…


```
"qid": "e0a9f1ef5fae44427b26",
"term": "Pear",
"description": "genus of plants",
"question": "Would a pear sink in water?",
"answer": false,
"facts": [
    "The density of a raw pear is about 0.59 g/cm^3.",
    "The density of water is about 1 g/cm^3.",
    "Objects only sink if they are denser than the surrounding fluid."
],
"decomposition": [
    "What is the density of a pear?",
    "What is the density of water?",
    "Is #1 greater than #2?"
],
```


Oh, they crowdsourced the data set from Mechanical Turk (a platform which pays random people in poor countries a few cents to do mindless tasks as quickly as possible). They were supposed to use only Wikipedia to generate Q/A pairs. But it

doesn't give pear density, so …



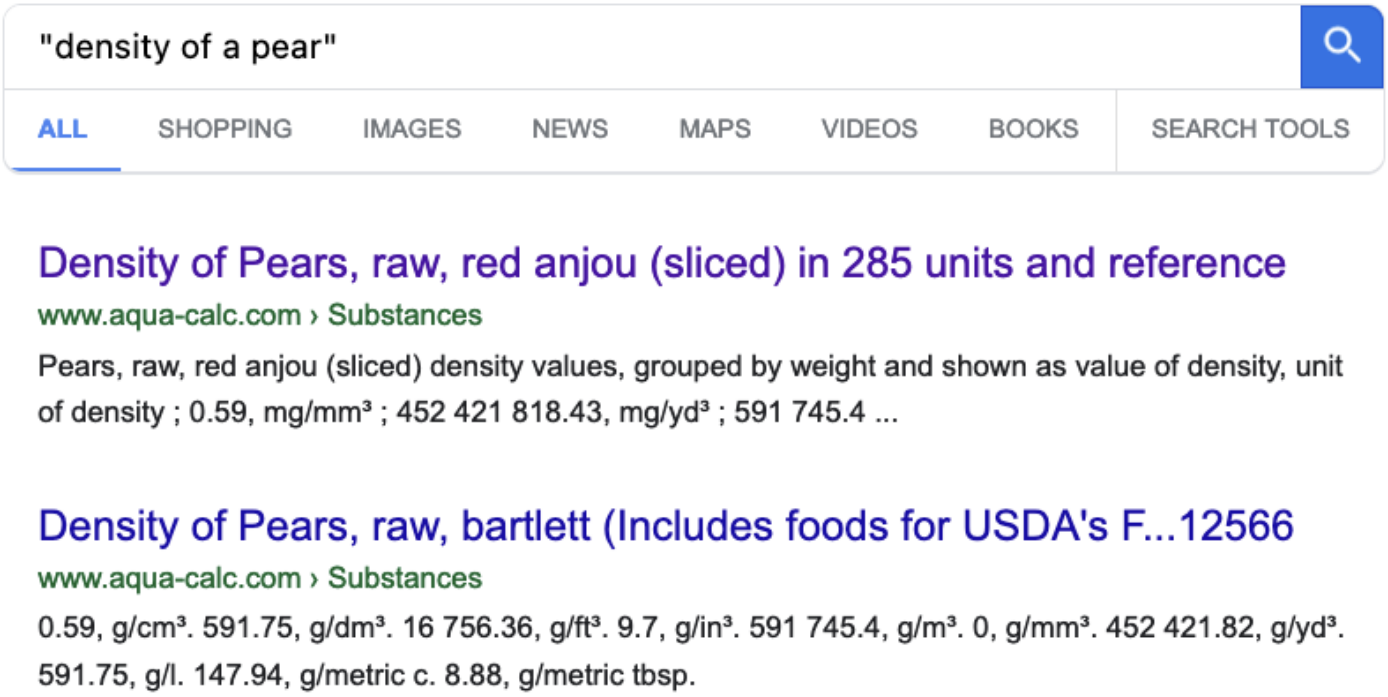
3 Data Collection Pipeline

Our goal is to establish a procedure for collecting strategy questions and their annotations at scale. To this end, we build a multi-step crowdsourcing[1] pipeline designed for encouraging worker creativity, while preventing biases in the data.

[1]We use Amazon Mechanical Turk as our framework.

The first hit in my web search for "density of a pear" explains that *sliced* pears have a density of 0.59, which presumably is where the Mechanical Turk worker got it.

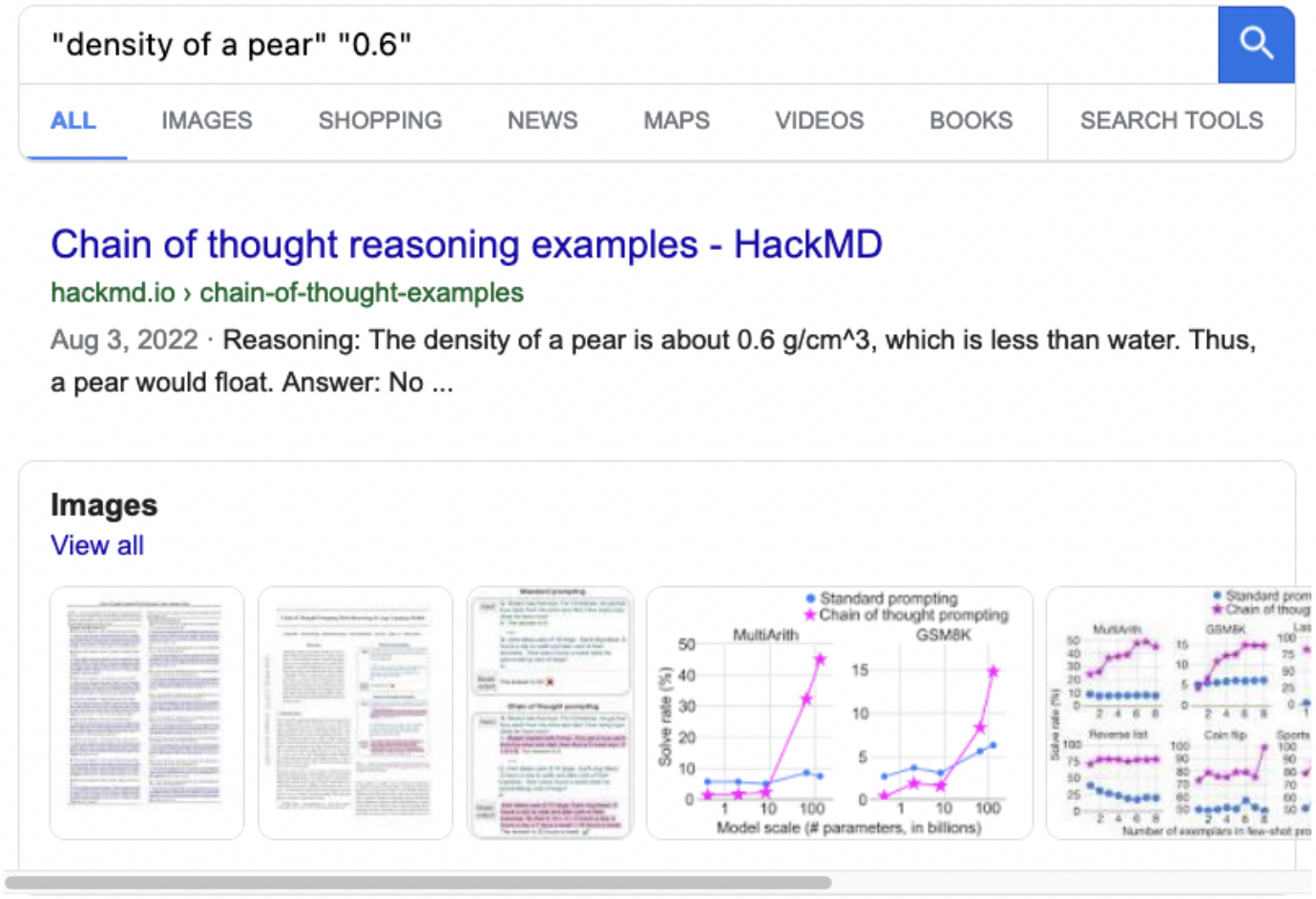
(No idea where this came from; someone weighted a package of sliced pairs that was mostly air I suppose.)



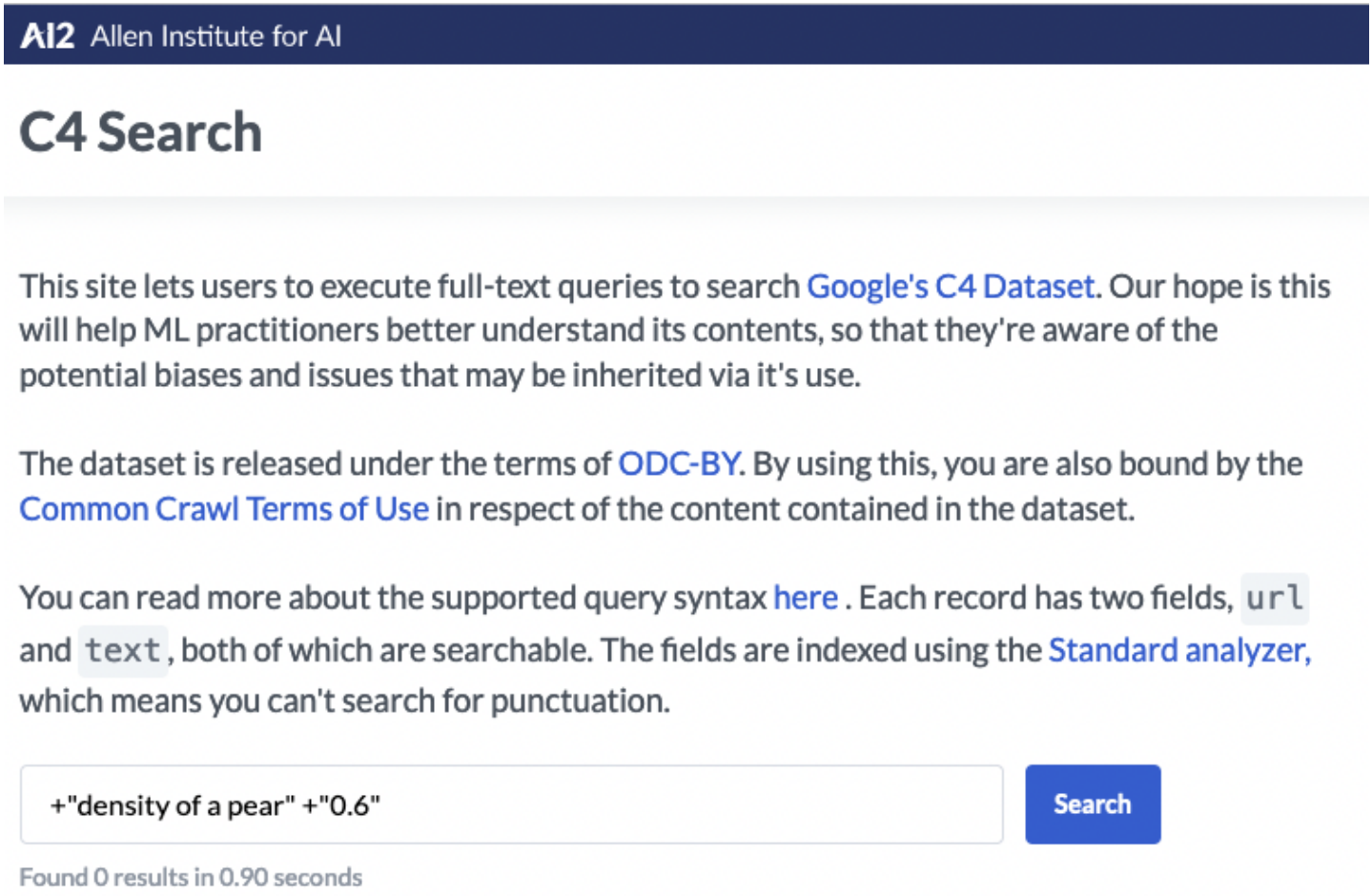
So where did the (wrong) answer in the Google AI blog come from? They *don't* say it was produced by an AI. Maybe it's just an illustrative example, and a human turned the StrategyQA json data into cleaner text, rounding 0.59 to 0.6.

(If you are at Google AI, I'd like to know!)

I can’t find any statement that the density of pears is 0.6 on the web, other than downstream from this “chain of thought” AI experiment. But, if the answer *was* AI-generated, maybe it came from its larger-than-Google text database, or maybe it knows how to round!



And, hmm, not in the C4 dataset either, which is a biggish chunk of what language models get trained on.



Anyway, before I got distracted by ¡SCIENCE!: if that answer *was* generated by software, how close is it to something in the training data?

If it includes e.g. “The density of a pea is about 0.9 g/cm^3, which is less than water. Thus, a pea would float”, I’d be unimpressed

There’s many similar explanations on the web. How different is the closest to the maybe-AI-generated answer? I haven’t located one very close (but I don’t have access to the full training set, and we don’t have good tools for this sort of analysis).

Results

Different fruits and vegetables will also float or sink depending on their density. In general, apples, bananas, lemons, oranges, pears, and zucchinis will float, while avocados, potatoes, and mangoes will sink. Others like turnips and sweet potatoes sometimes sink and sometimes float.

Why?

Whether a fruit or vegetable sinks or floats has a lot to do with its density. What is density? It's how heavy an object is compared to its volume. Imagine lifting a pillowcase full of feathers. Now imagine lifting a pillowcase full of apples. Which one would be heavier? They are both the same size, but since apples are denser than feathers, the one with the apples in it would be heavier.

In this experiment, you were trying to find out what vegetables and fruits are the lightweights of the plant world. Fruits and vegetables are dense for different reasons. Some are like pumpkins, with thick skins. Others are like avocados, with a big pit. Others have more air inside them. The shape or age of a fruit can also impact whether it sinks or floats.

Why do objects float?

Each fruit or vegetable has a volume. The volume is the length multiplied by the width multiplied by the height. Water has volume, too. When you put an object into water, it pushes the water aside. The water tries to go back down into the hole.

Who will make it to the bottom, and which one will float on top? Water and apples don't arm wrestle to decide. Instead, it's based on the properties of matter. Whatever is less dense will float. Since apples are usually less dense than water, they will float on top of the water. If you place an avocado into the water, it is denser than water, so it will sink. The water is less dense, so it will "float" on top of the avocado. If your container is completely full when you put the avocado in, the water will spill over the top.

If the answers in the Google write-ups were just illustrative examples written by a human, it would be good for the team to clarify that. I am not the only one who was wowed by them, believing them to have been machine-generated.

Here is GPT-3: "Pears are less dense than water, so they would float."

Q: Yes or no: Would a pear sink in water?

A: Pears are less dense than water, so they would float.

Interestingly, this answer is found virtually verbatim on the web in lots of places, except with apples! It's a really easy answer to generate, which is making me think some more that the 0.6 gm/cm^3 answer was written by a human, based on the human-generated StrategyQA answer.

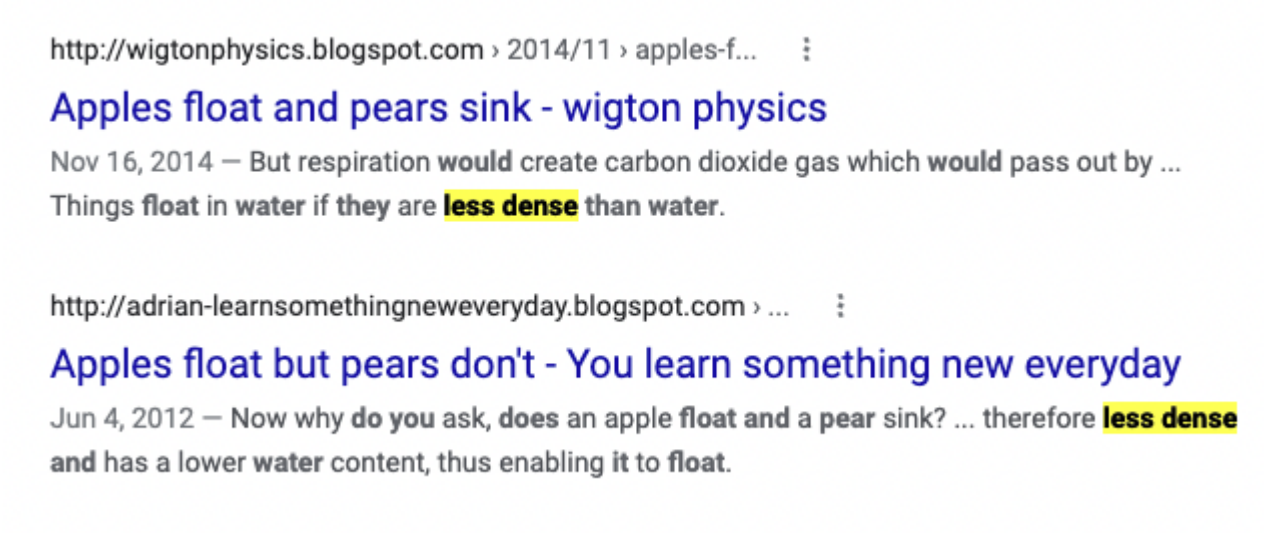
https://www.quora.com › How-do-I-explain-to-my-daugh...

How to explain to my daughter why apples float in water but ...

If an item is **more dense than water it will** sink, it is **less dense it will float**. Apples are **less dense than water, so** they **float**!

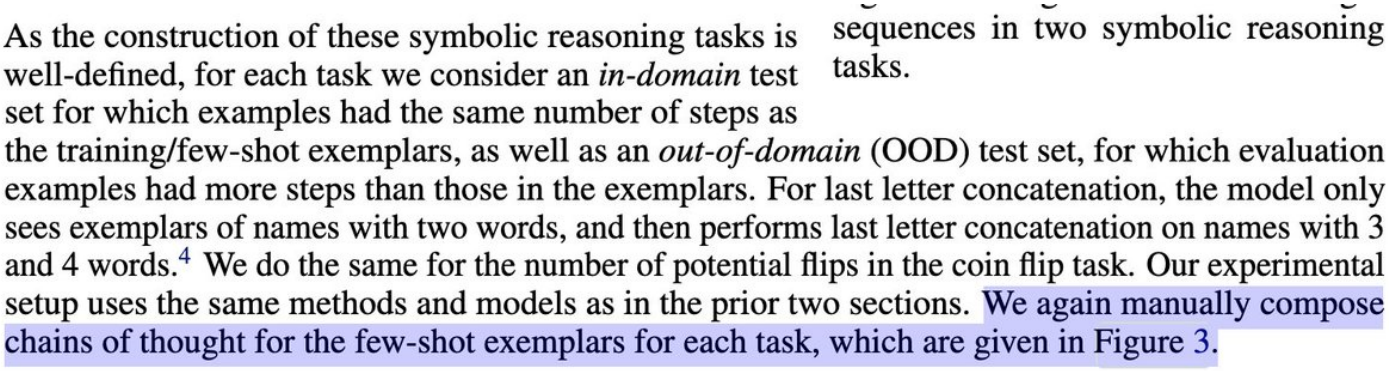
10 answers · 23 votes: Assuming your daughter is young I'd consider saying: Things that are pa...

Meanwhile, there's tons of things on the web that say that pears sink.



■ This isn't a big deal and I have spent way too much time on it. However, I can report that the pear was delicious, so it wasn't a complete waste!

OK, reading the actual paper (rather than the blog post) carefully makes it fairly clear that this answer *was* written manually, as an example for zero-shot learning. I sent myself off on a wild pear chase, apparently!
https://t.co/ERYmqWyQN8



sequences in two symbolic reasoning tasks.

As the construction of these symbolic reasoning tasks is well-defined, for each task we consider an *in-domain* test set for which examples had the same number of steps as the training/few-shot exemplars, as well as an *out-of-domain* (OOD) test set, for which evaluation examples had more steps than those in the exemplars. For last letter concatenation, the model only sees exemplars of names with two words, and then performs last letter concatenation on names with 3 and 4 words.[4] We do the same for the number of potential flips in the coin flip task. Our experimental setup uses the same methods and models as in the prior two sections. We again manually compose chains of thought for the few-shot exemplars for each task, which are given in Figure 3.