(J) Strawbery Shortcake (1/2) [5 Points]

Real-world AI language models sometimes confidently state things that are completely made up; this is often called a **hallucination**. Recently, NACLOLabs has hired you to help understand why its new chatbot (an artificial intelligence system designed to have conversations), ParrotBot, has been hallucinating. Instead of understanding facts or meaning, ParrotBot only cares about patterns. Here is ParrotBot's single rule for generating text:

> *Look at the last word I just wrote, look through my training data to see what word follows it most frequently, and output that word.*

Here, *training data* means the collection of example text that ParrotBot was trained on. For example, consider ParrotBot trained on the training data 1a–1c shown on the right. Given an initial input `poison` and asking it to output 5 more words after, ParrotBot's response is generated using its method as follows:

Training data 1:
1a. poison can be bad for you
1b. vitamins can be good for you
1c. exercise can be good for people

Initial input → `poison`

1. After `poison`, only `can` appears (1x). → `can`
2. After `can`, only `be` appears (3x). → `be`
3. After `be`, `good` (2x) beats `bad` (1x). → `good`
4. After `good`, only `for` appears (2x). → `for`
5. After `for`, `you` (2x) beats `people` (1x). → `you`

Therefore, ParrotBot's full message is "`poison can be good for you`," which wasn't in the training data at all! This shows how our simple language model can create factually untrue sentences even when only provided true ones.

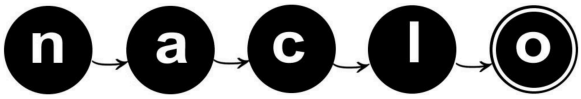
J1. NACLOLabs creates a new chatbot called ParrotBot$_2$, which is trained on *only* the data 2a–2e below. Given an initial input `small` and asking it to output 4 more words after, write down ParrotBot$_2$'s full message:

Training data 2:
2a. small elephants eat small strawberries
2b. small elephants are not spies
2c. some spies are alleged spies
2d. alleged spies are very sneaky
2e. no elephants are alleged blueberries

`small` | | | | |

Modern AI language models are more sophisticated than ParrotBot, but they still sometimes make mistakes. One of the most widely known hallucinations in modern AI language models involves letter-based reasoning. For example, when asked how many times the letter *r* appears in the word *strawberry*, language models often give incorrect answers such as *2*.

The mistake most likely comes from how the model represents text. Instead of working letter-by-letter, it usually breaks text into pieces called **tokens** and then works with those tokens. A token is a piece of text the model processes as one unit—sometimes a single letter, sometimes a fragment of a word, and sometimes a whole word. As such, it can struggle with tasks that require careful reasoning at the individual-letter level.



(J) Strawbery Shortcake (2/2)

Because tokens are built from frequent patterns in text, they often look like meaningful pieces of words, but they do not always line up with meaning. For example, one possible tokenization of the sentence *She read the bread is heavy* is [*S*][*he*] [*read*] [*t*][*he*] [*b*][*read*] [*is*] [*he*][*a*][*vy*]. Notably, the sequence *read* inside *bread* is just part of the word, not a meaningful piece on its own, and the sequence *he* inside *the* does not mean the pronoun *he*.

Below is a message from an AI language model, responding to a question asking it to count how many letters are common between certain pairs of words. However, instead of reasoning at the level of individual letters, it used its internal token representation; that is, it counted how many *tokens* are exactly the same, and then reported that number as though it were the number of letters in common.

The six words given to the AI have been replaced with numbered stickers ① – ⑥; the same word is always covered by the same sticker.

> ① and ② have two letters in common.
> ① and ③ have two letters in common.
> ④ and ⑥ have two letters in common.
> ① and ④ have one letter in common.
> ② and ⑤ have one letter in common.
> ④ and ⑤ have one letter in common.
> ① and ⑥ have no letters in common.
> ② and ④ have no letters in common.
> ③ and ⑤ have no letters in common.
>
> None of these words have any repeating letters, so I didn't need to worry about how to count duplicates. And interestingly enough, all the words are four letters long!

Here are the words ① – ⑥, in arbitrary order:

A. *changing* B. *coaches* C. *excessive* D. *proceeding* E. *processes* F. *started*

J2. Match ① – ⑥ to their appropriate words A–F.
Hint: It is known that the model tokenizes the word *coaches* as [*co*][*a*][*ch*][*es*].

① ☐ ② ☐ ③ ☐ ④ ☐ ⑤ ☐ ⑥ ☐

J3. How does the model tokenize the word *started*?

started ☐