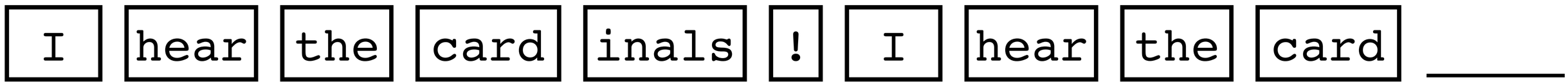
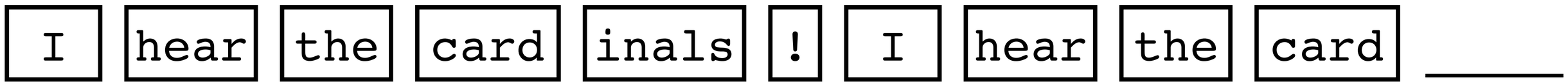
Complete the sequence: what comes next?

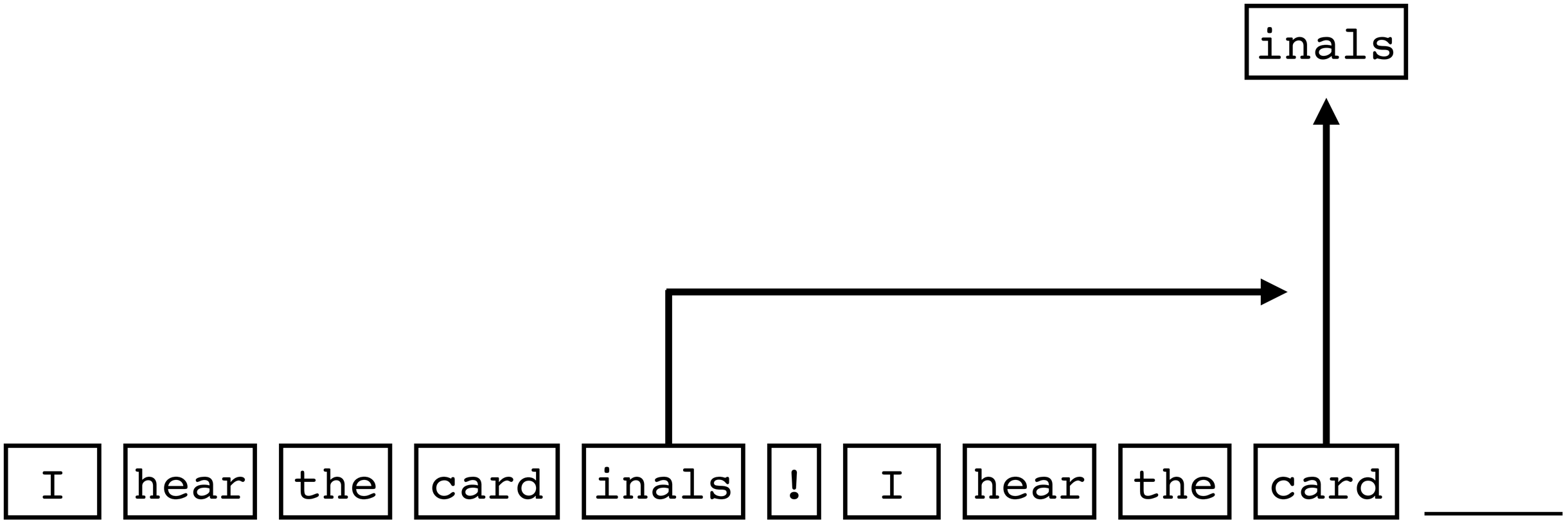


Complete the sequence: what comes next?

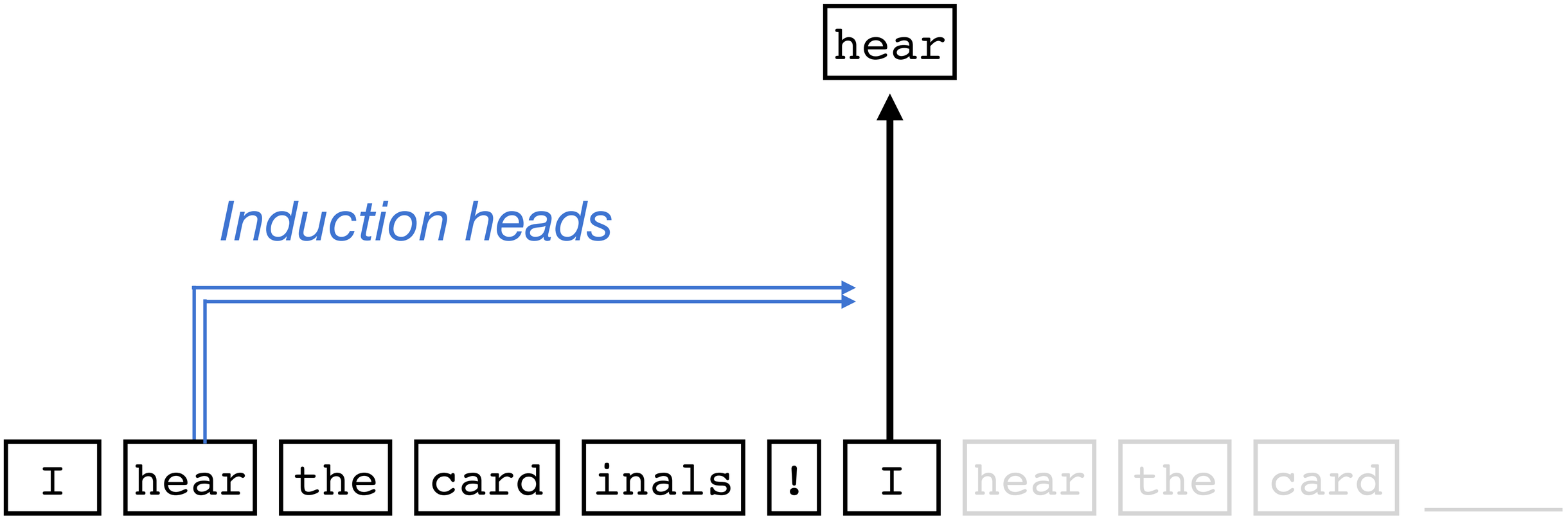
Probably “inals”, right?



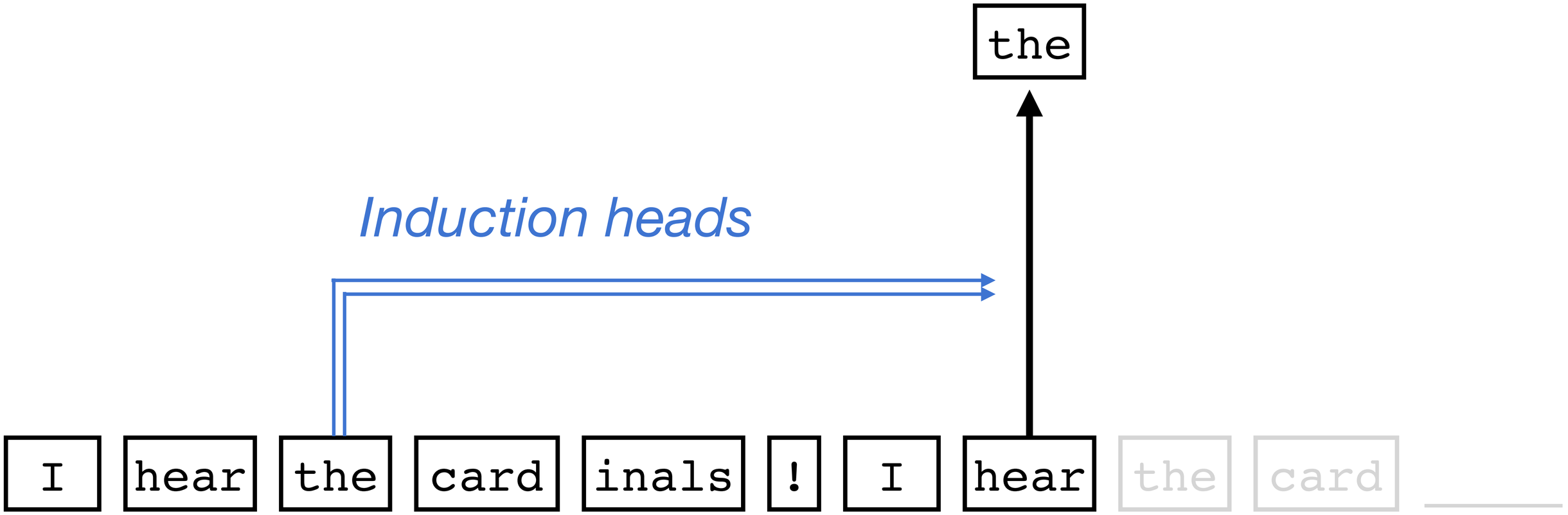
LLMs are really good at completing these sequences.



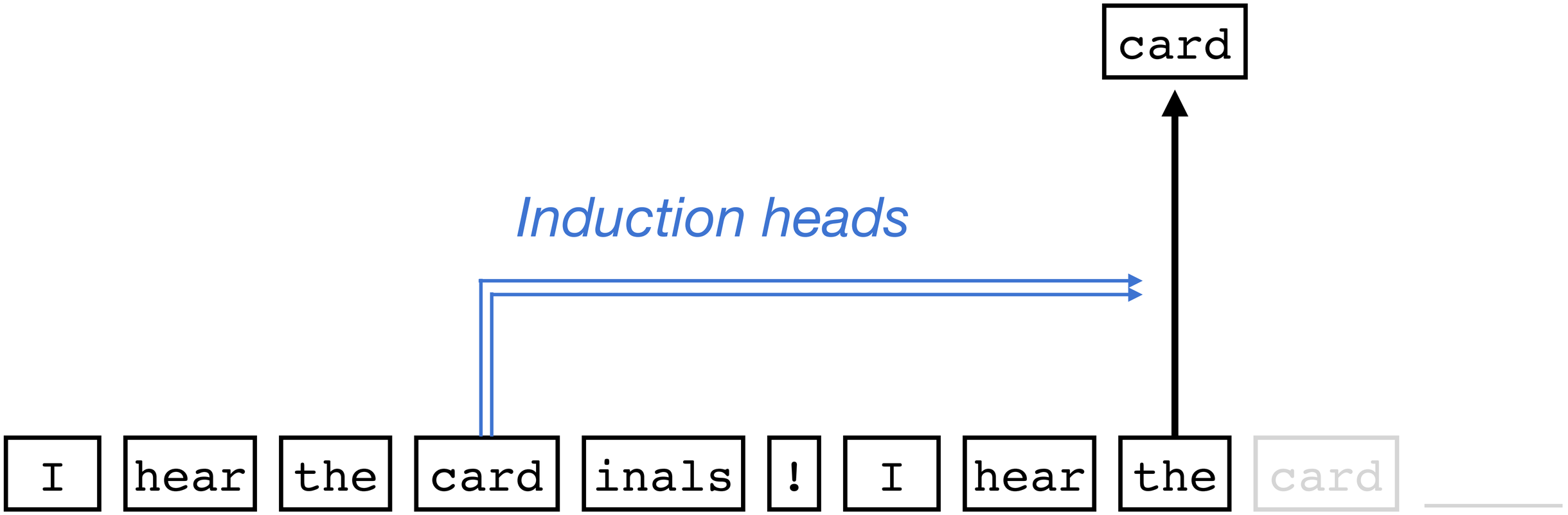
They can use induction heads at every token position to copy the whole sequence.



They can use induction heads at every token position to copy the whole sequence.

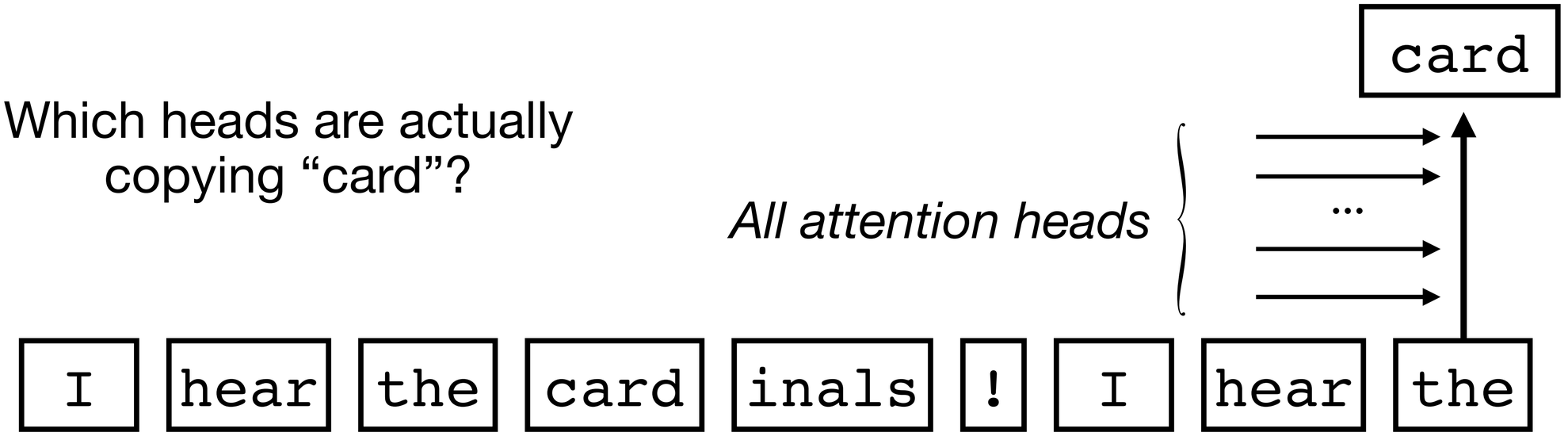


They can use induction heads at every token position to copy the whole sequence.



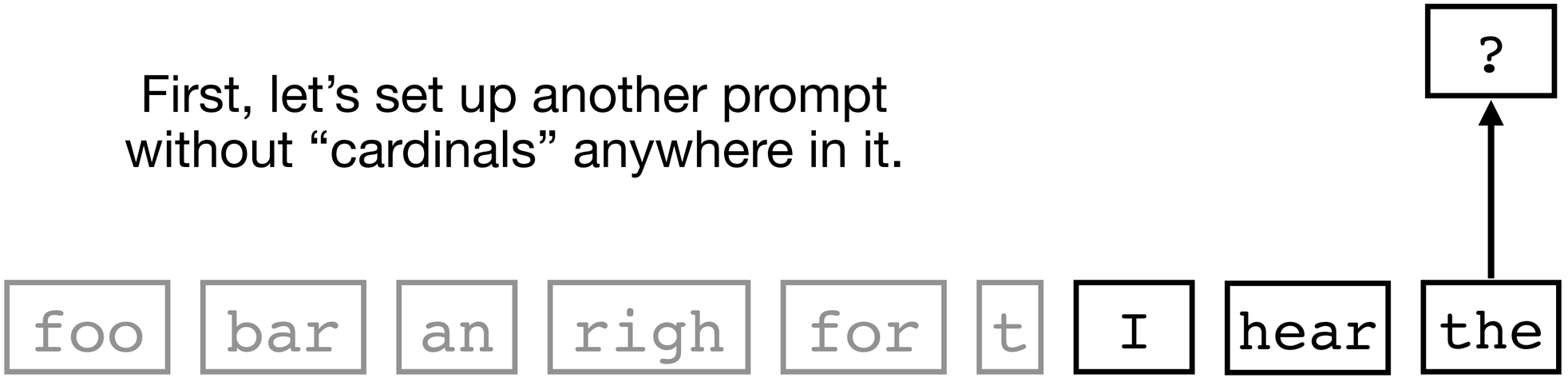
How do we find attention heads that copy?

Which heads are actually copying “card”?

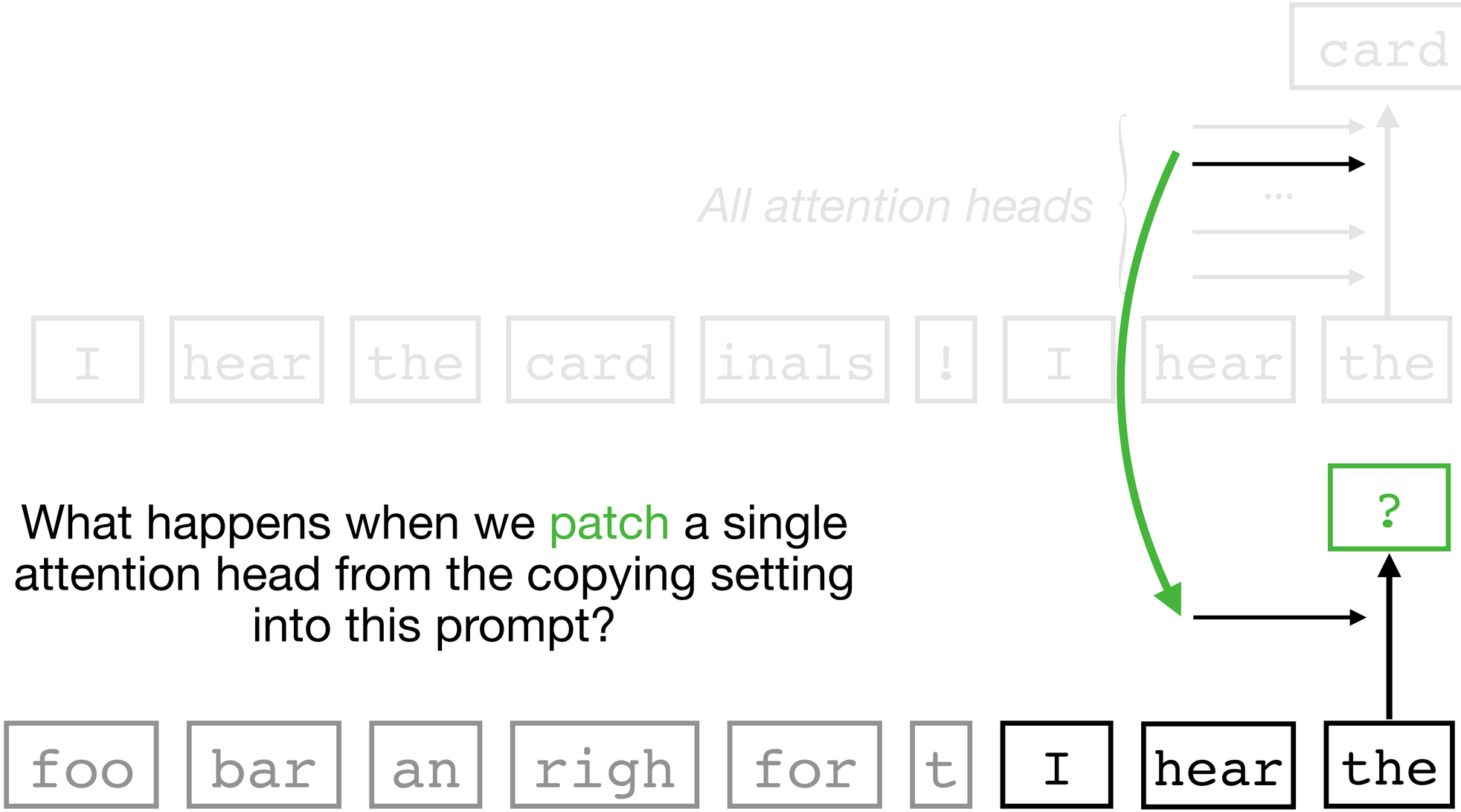


How do we find attention heads that copy?

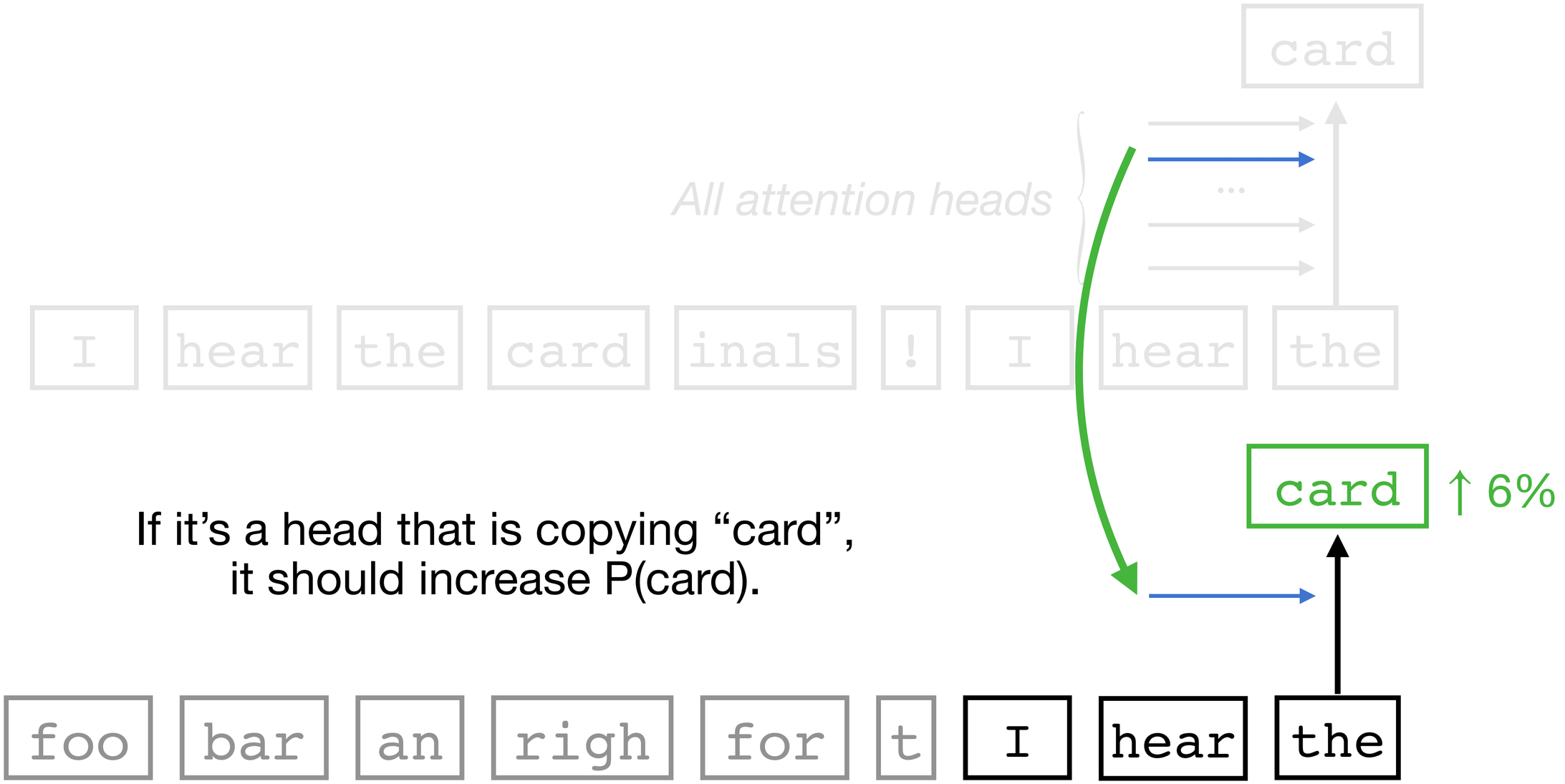
First, let's set up another prompt
without "cardinals" anywhere in it.



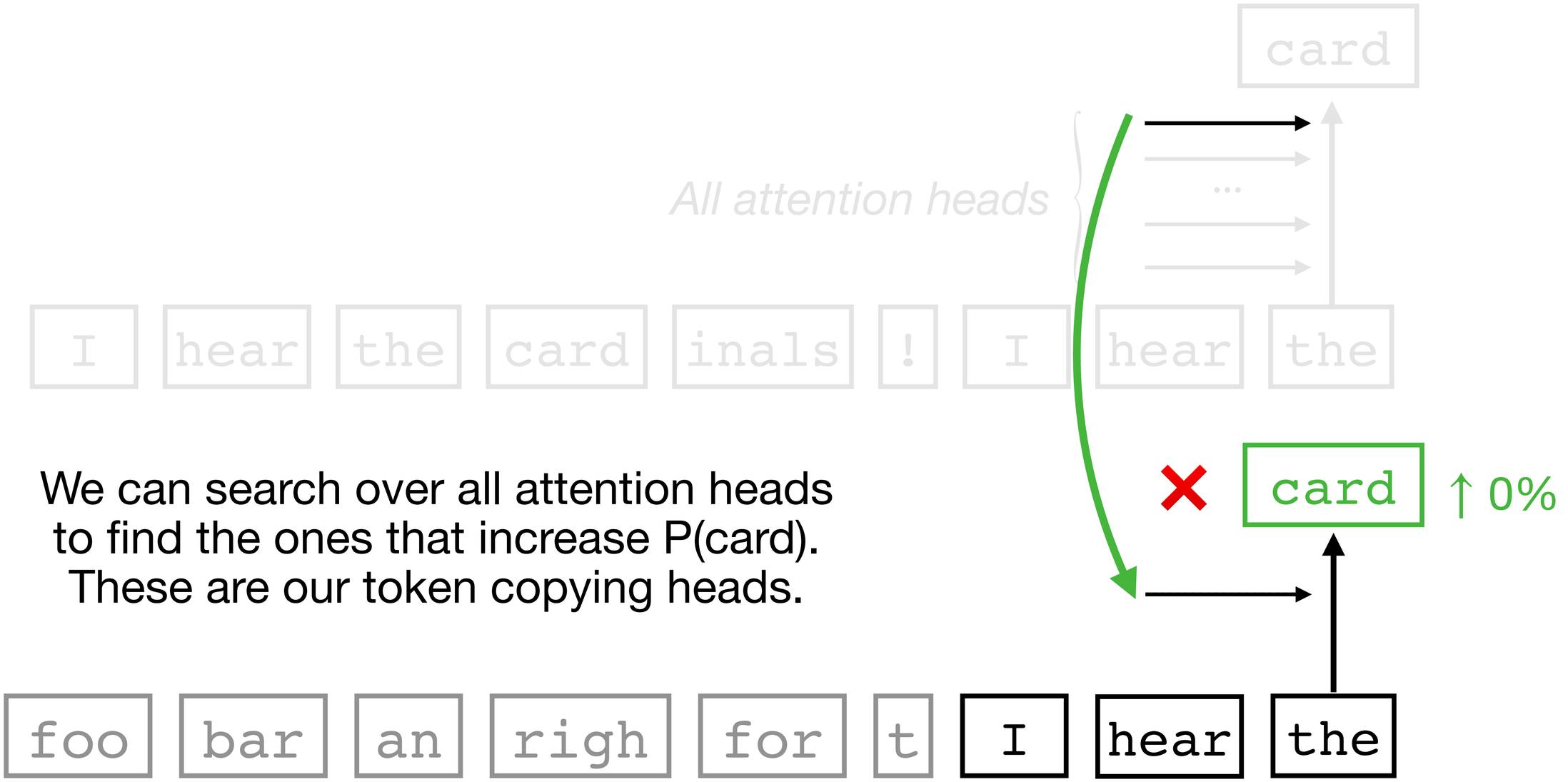
How do we find attention heads that copy?



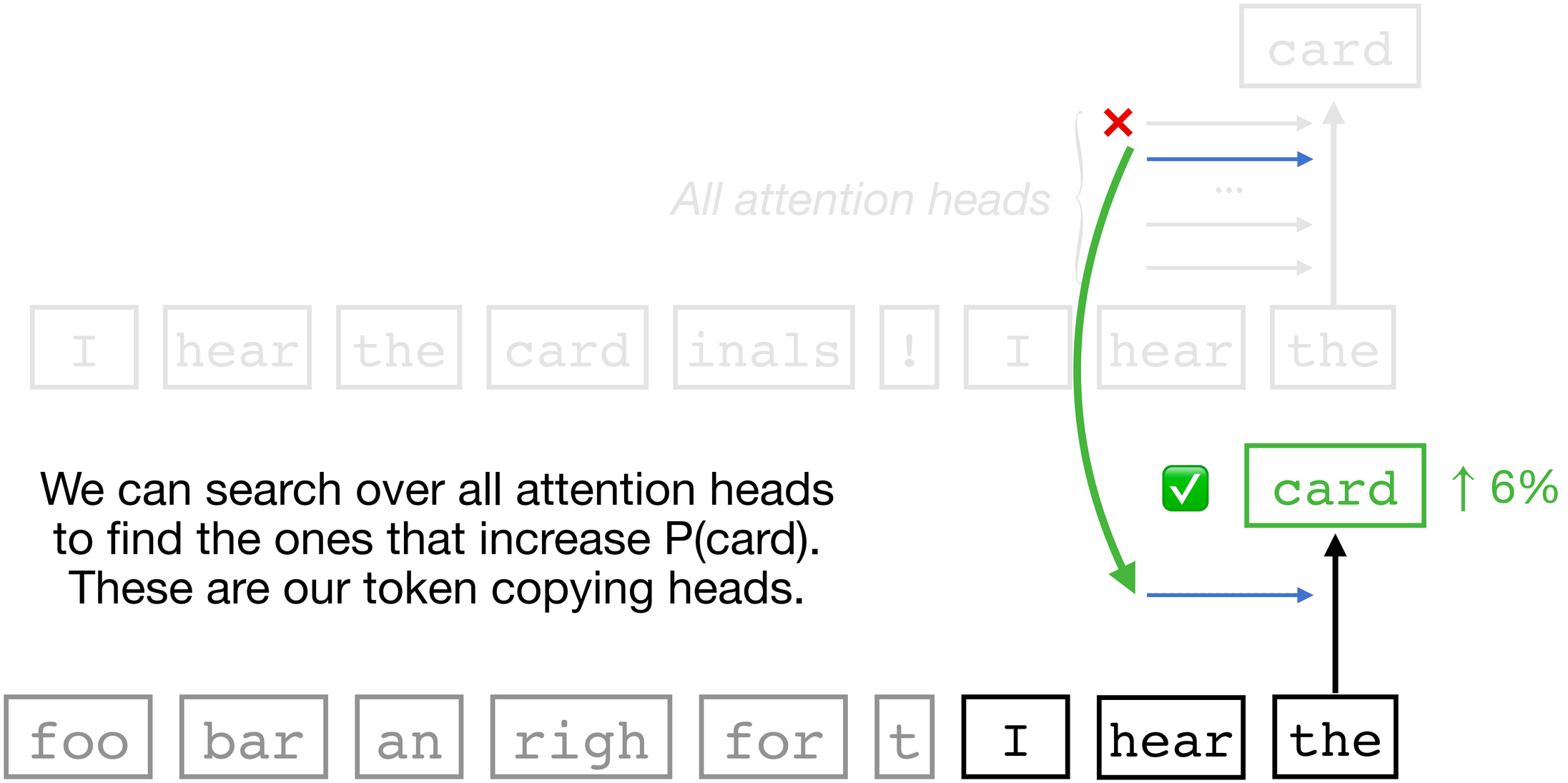
How do we find attention heads that copy?



How do we find attention heads that copy?

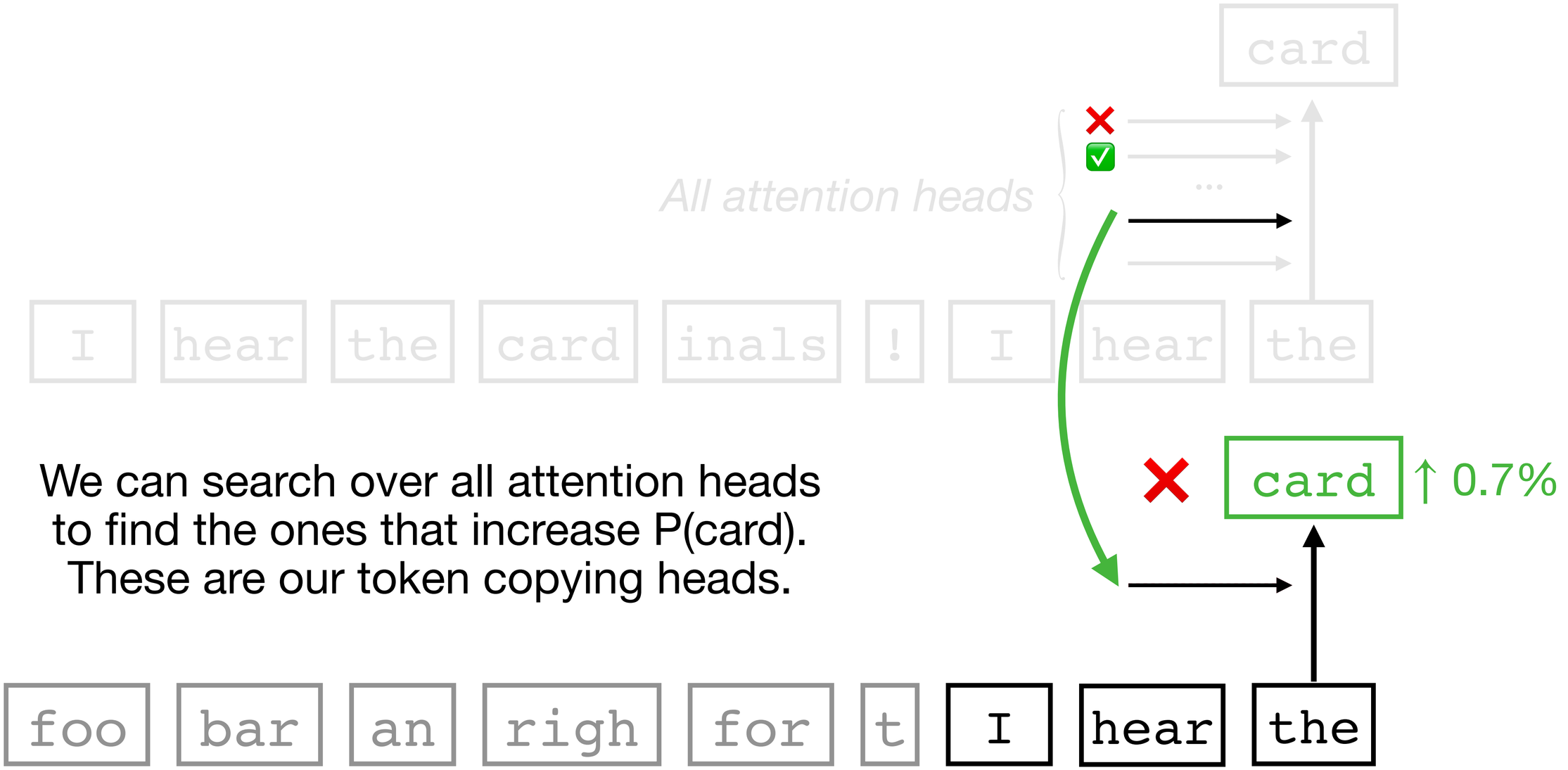


How do we find attention heads that copy?



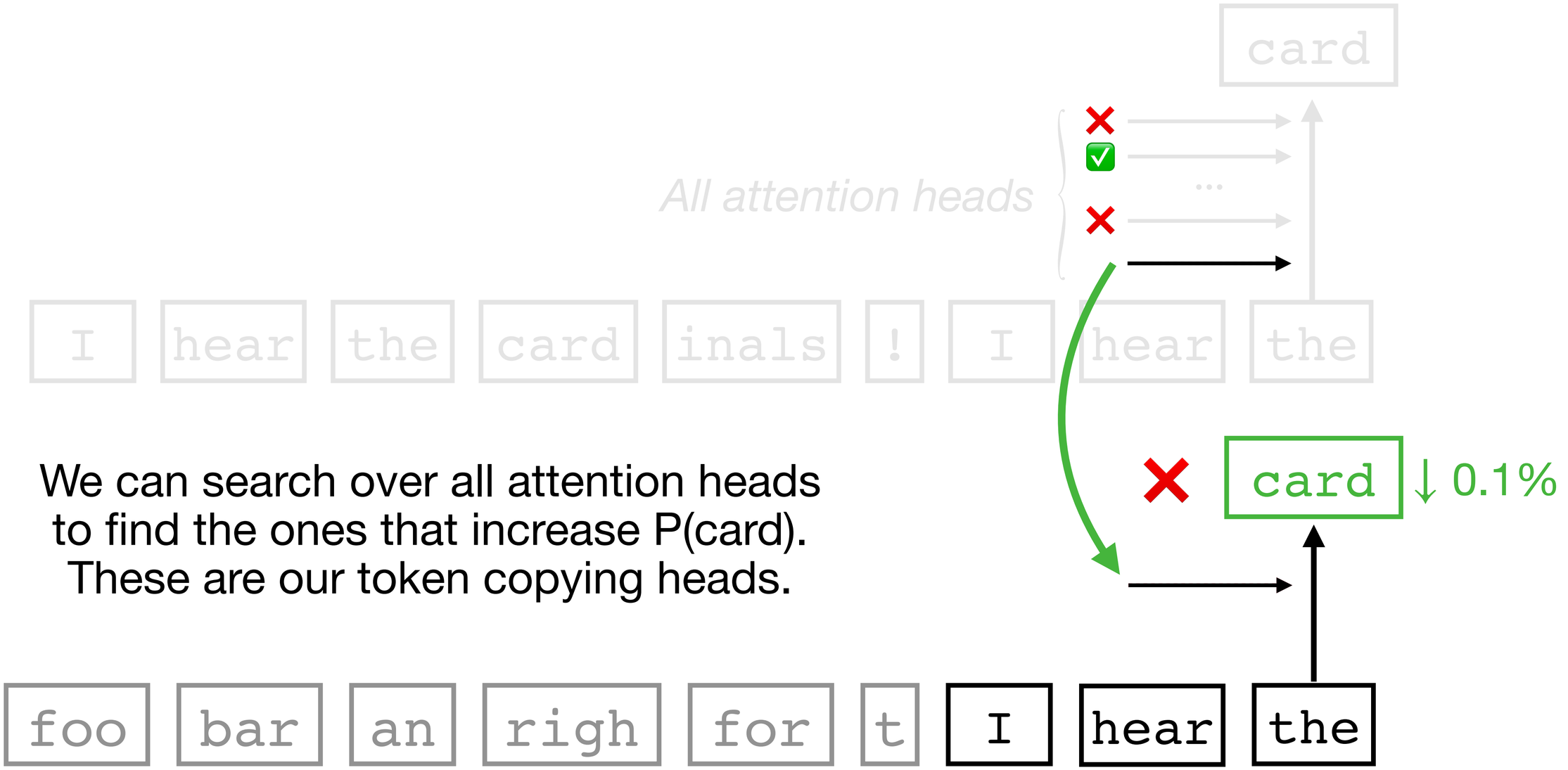
We can search over all attention heads
to find the ones that increase P(card).
These are our token copying heads.

How do we find attention heads that copy?



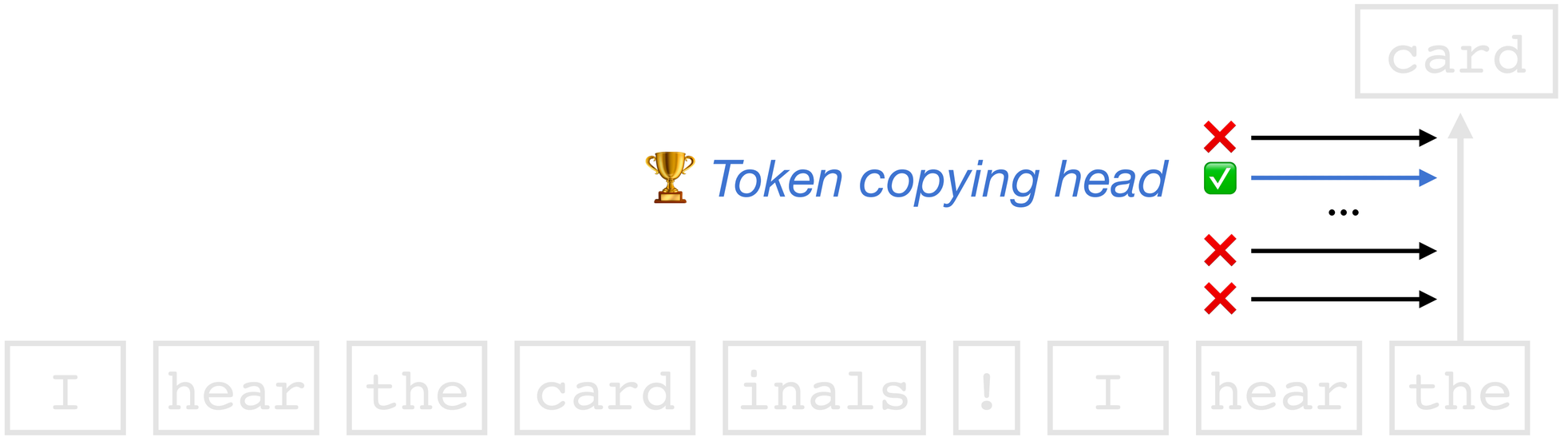
We can search over all attention heads
to find the ones that increase P(card).
These are our token copying heads.

How do we find attention heads that copy?

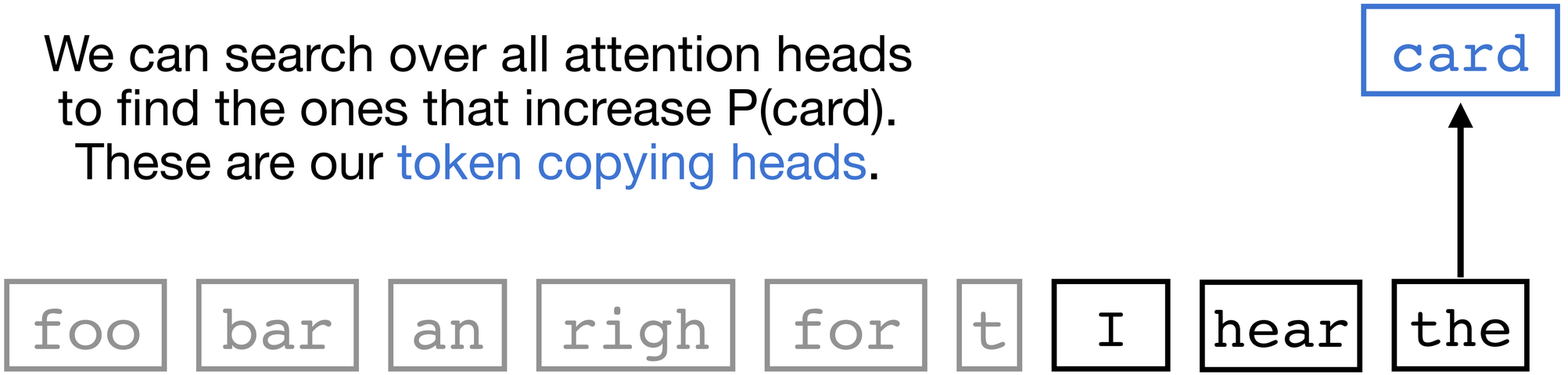


We can search over all attention heads
to find the ones that increase P(card).
These are our token copying heads.

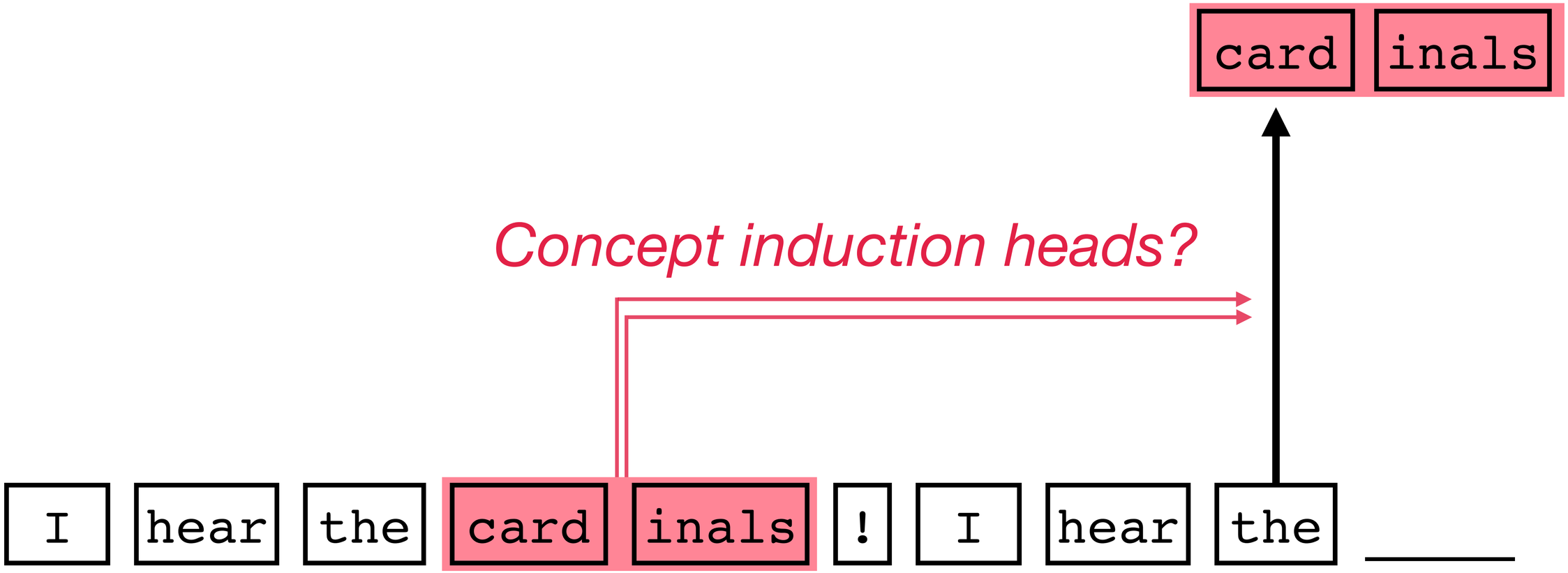
How do we find attention heads that copy?



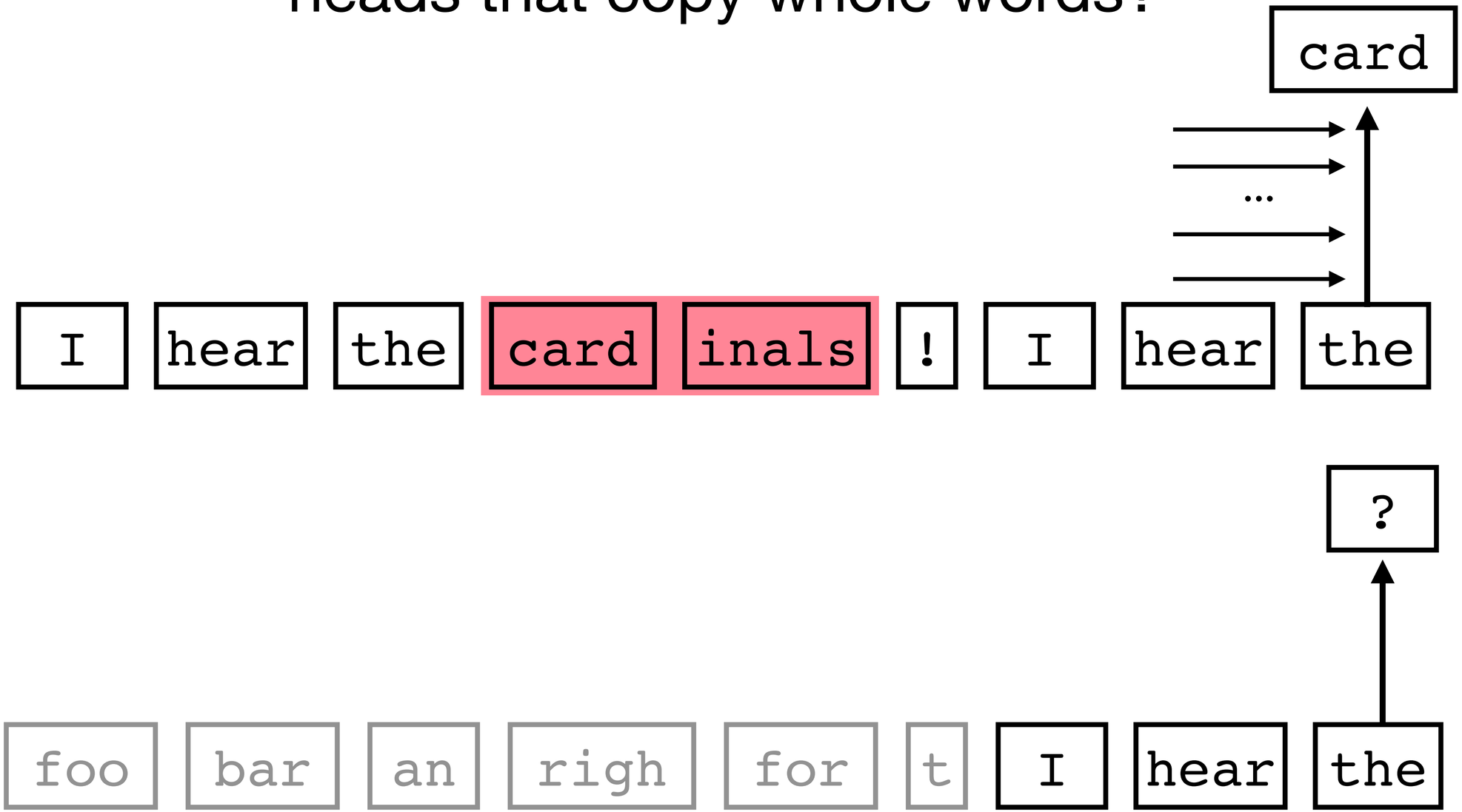
We can search over all attention heads
to find the ones that increase P(card).
These are our token copying heads.



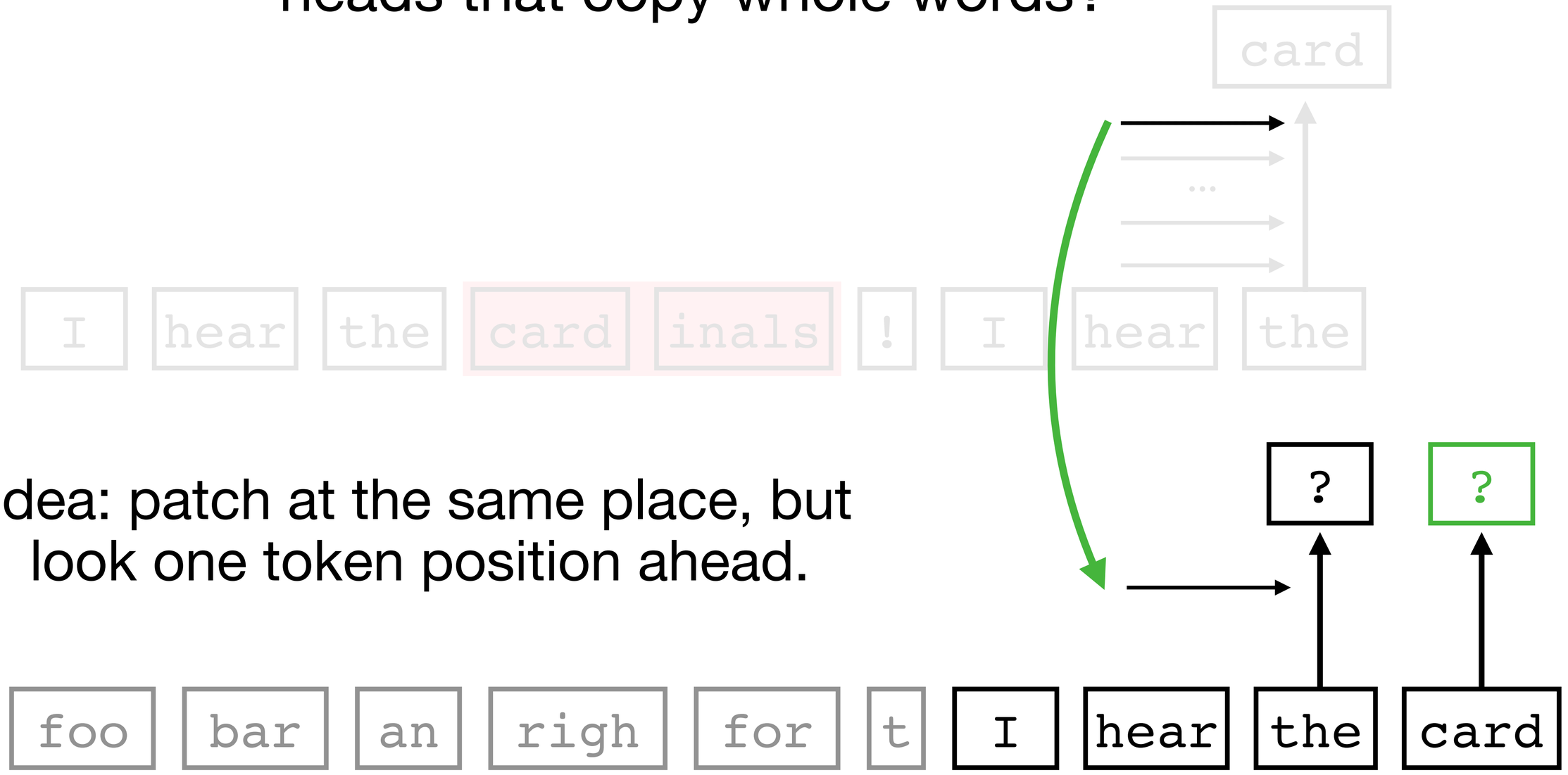
But what if, instead of copying one token at a time, the LLM copies entire *words* at a time?



Using patching, how could we find heads that copy whole words?

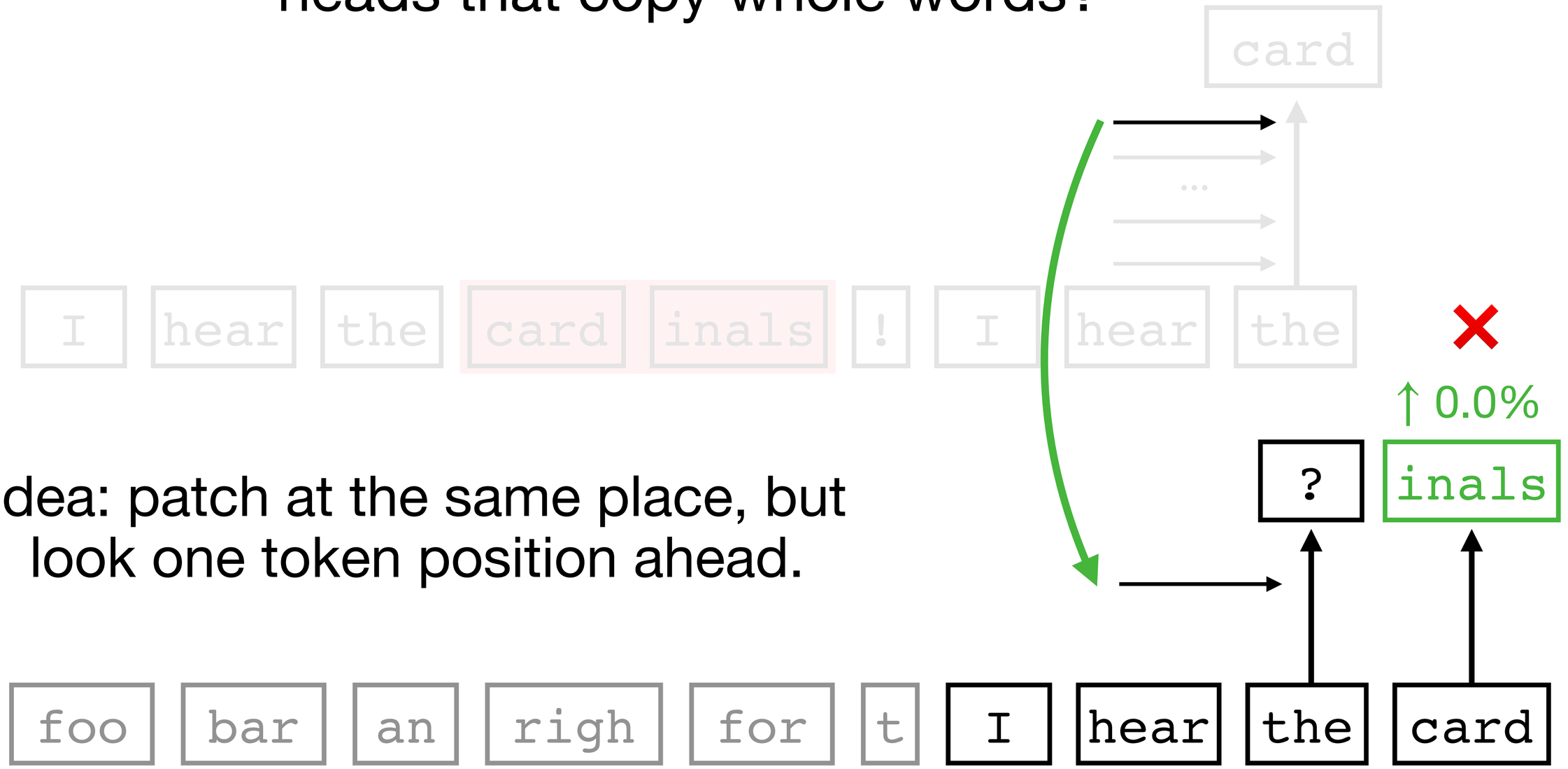


Using patching, how could we find heads that copy whole words?



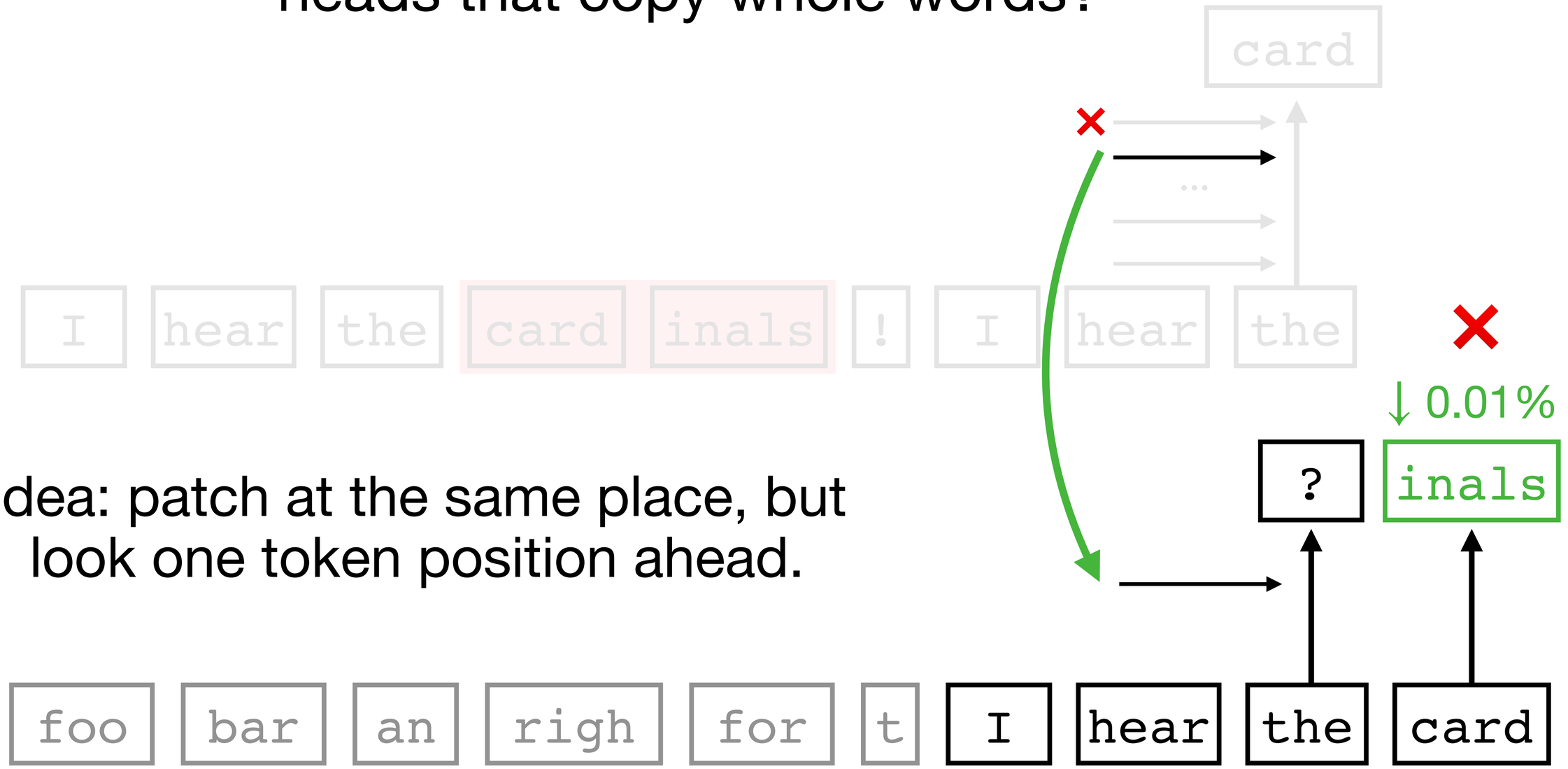
Idea: patch at the same place, but look one token position ahead.

Using patching, how could we find heads that copy whole words?



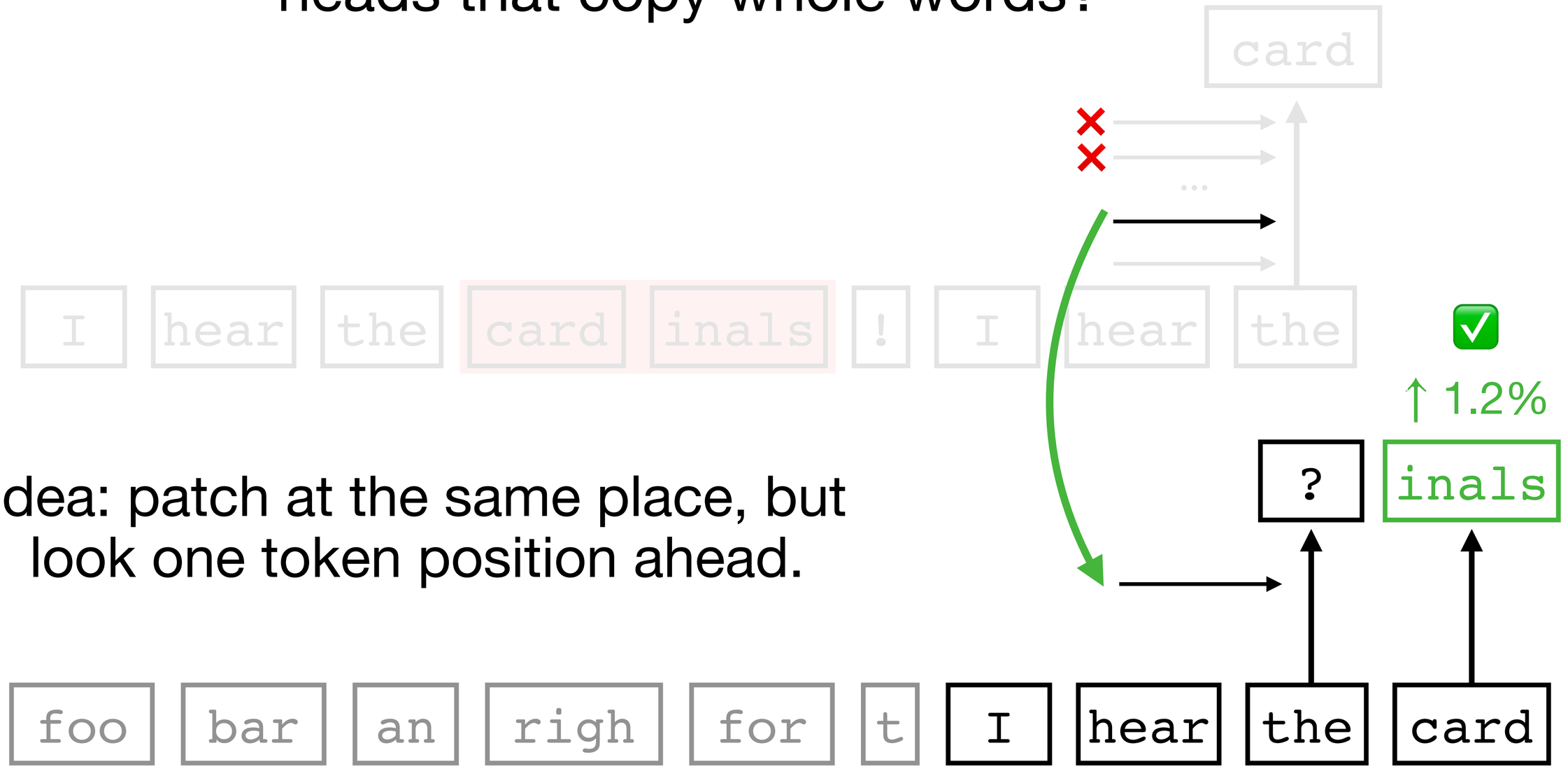
Idea: patch at the same place, but look one token position ahead.

Using patching, how could we find heads that copy whole words?



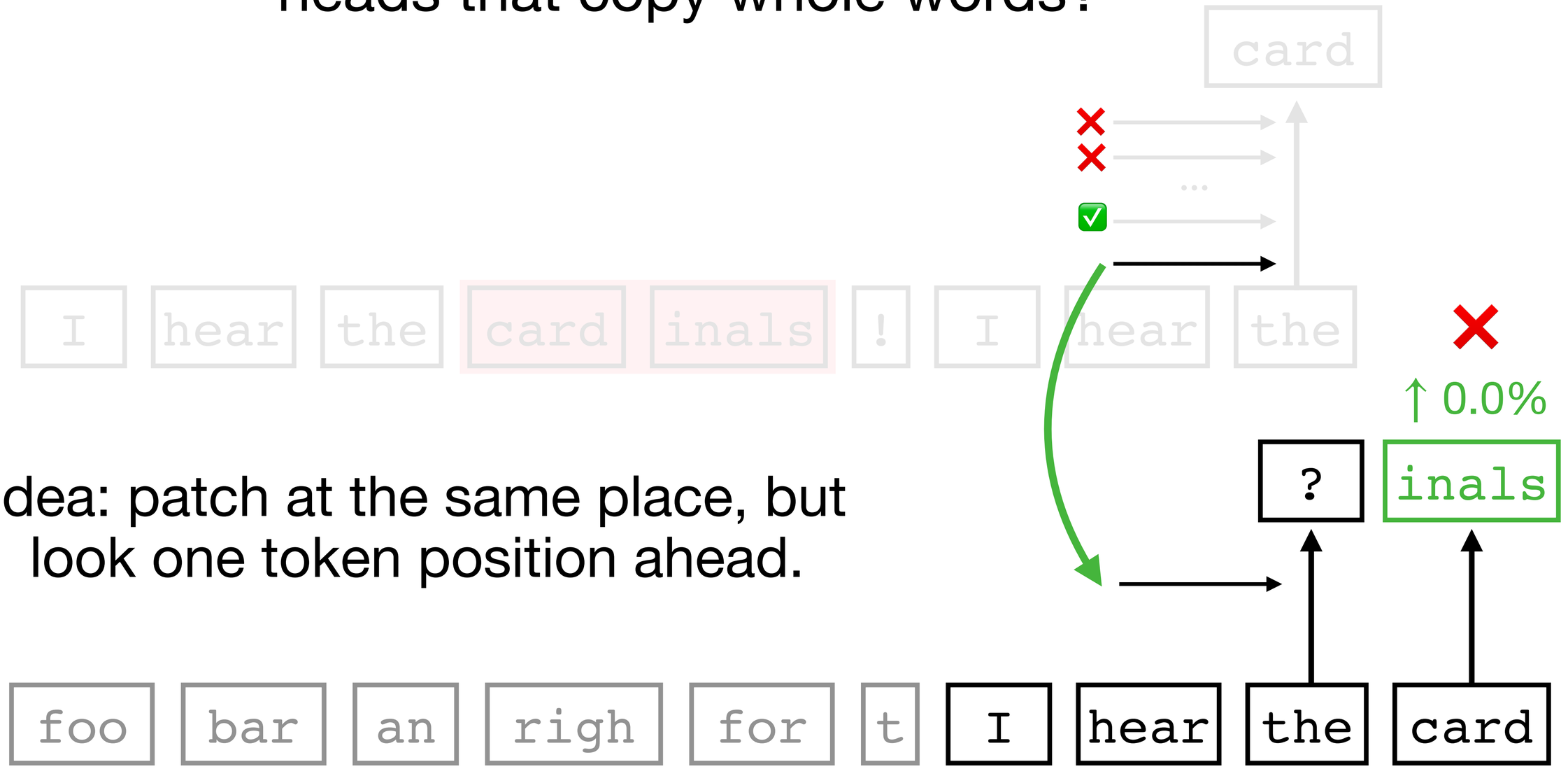
Idea: patch at the same place, but look one token position ahead.

Using patching, how could we find heads that copy whole words?



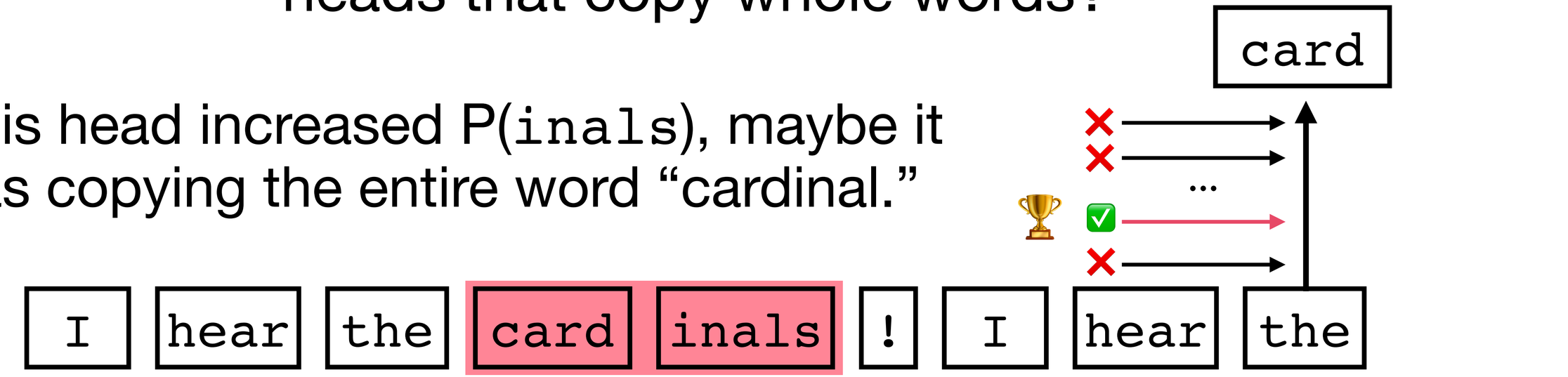
Idea: patch at the same place, but look one token position ahead.

Using patching, how could we find heads that copy whole words?

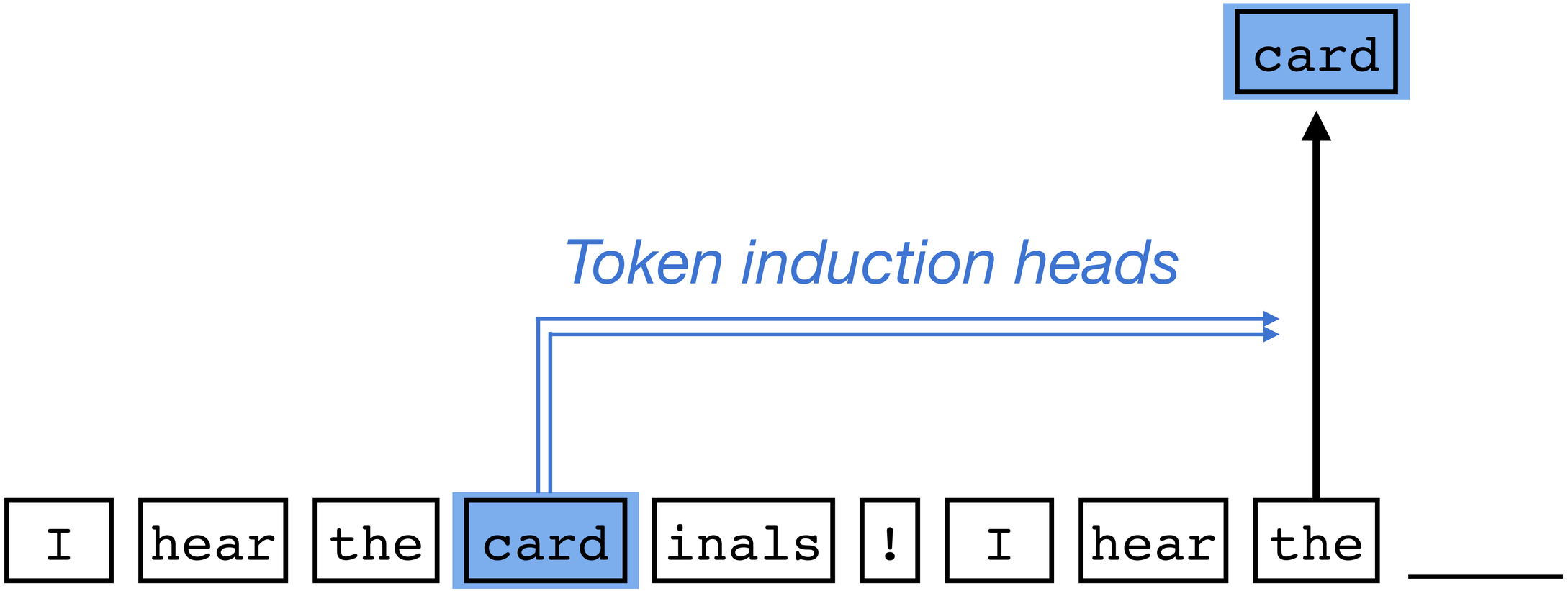


Using patching, how could we find heads that copy whole words?

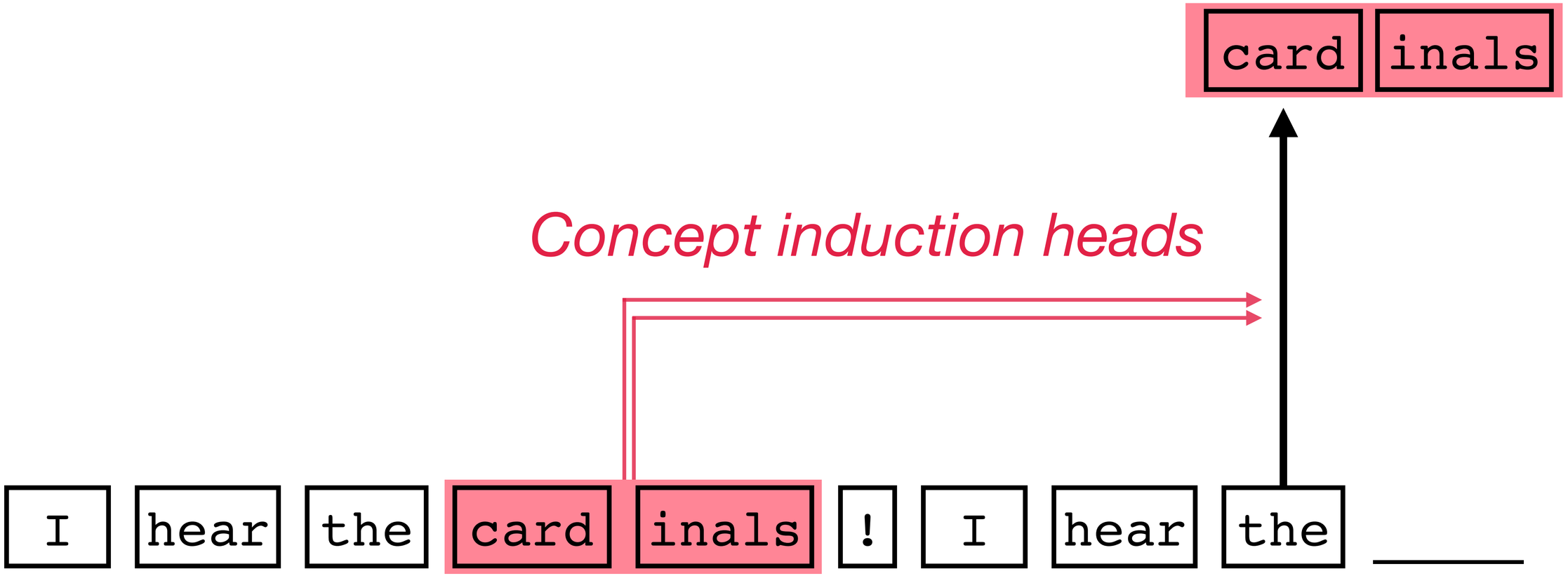
If this head increased P(`inals`), maybe it was copying the entire word “cardinal.”



So now we have one set of heads that increases P(`card`) at the next token…

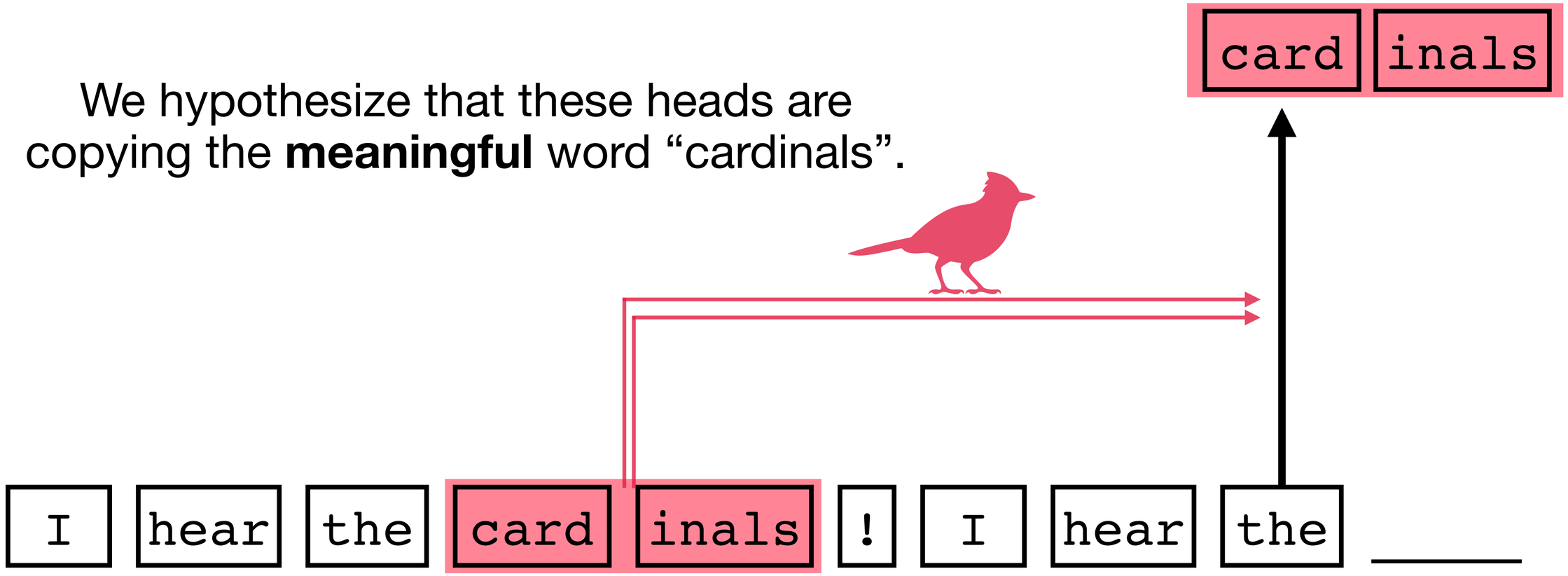


And another, separate set of heads that increases P(`inals`) at the next-next token.

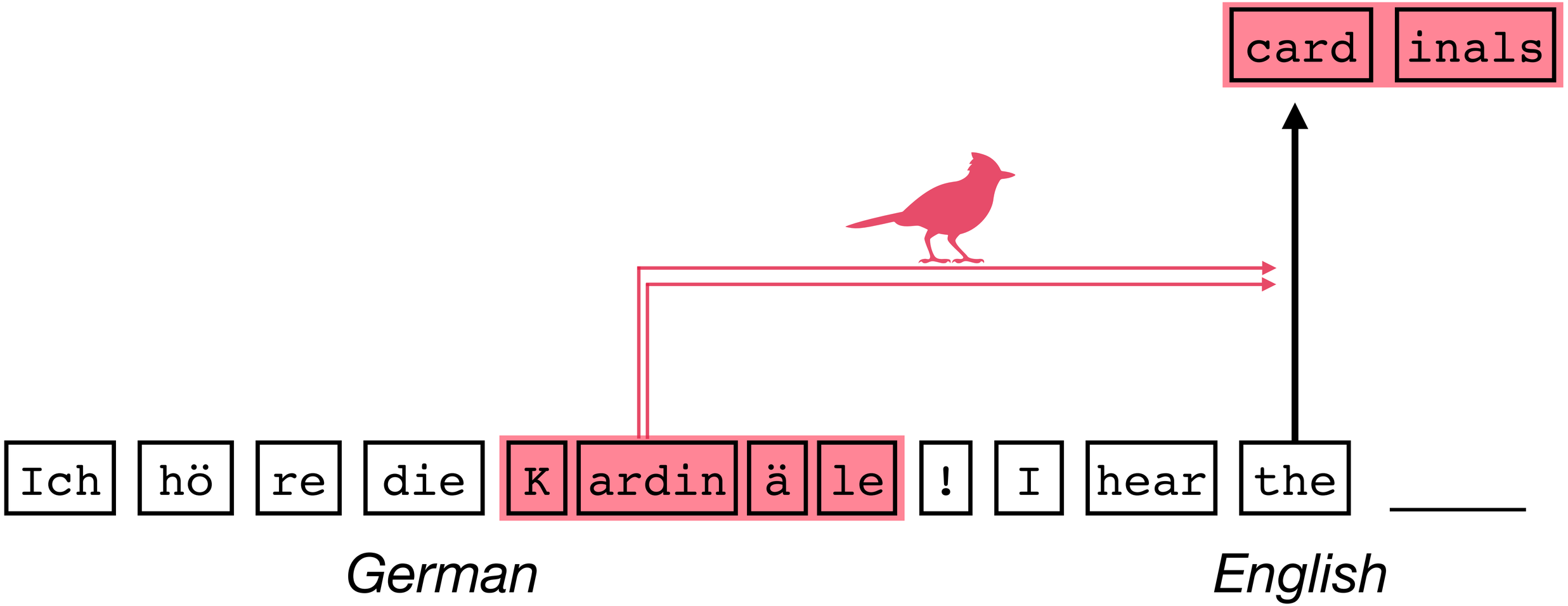


And another, separate set of heads that increases P(`inals`) at the next-next token.

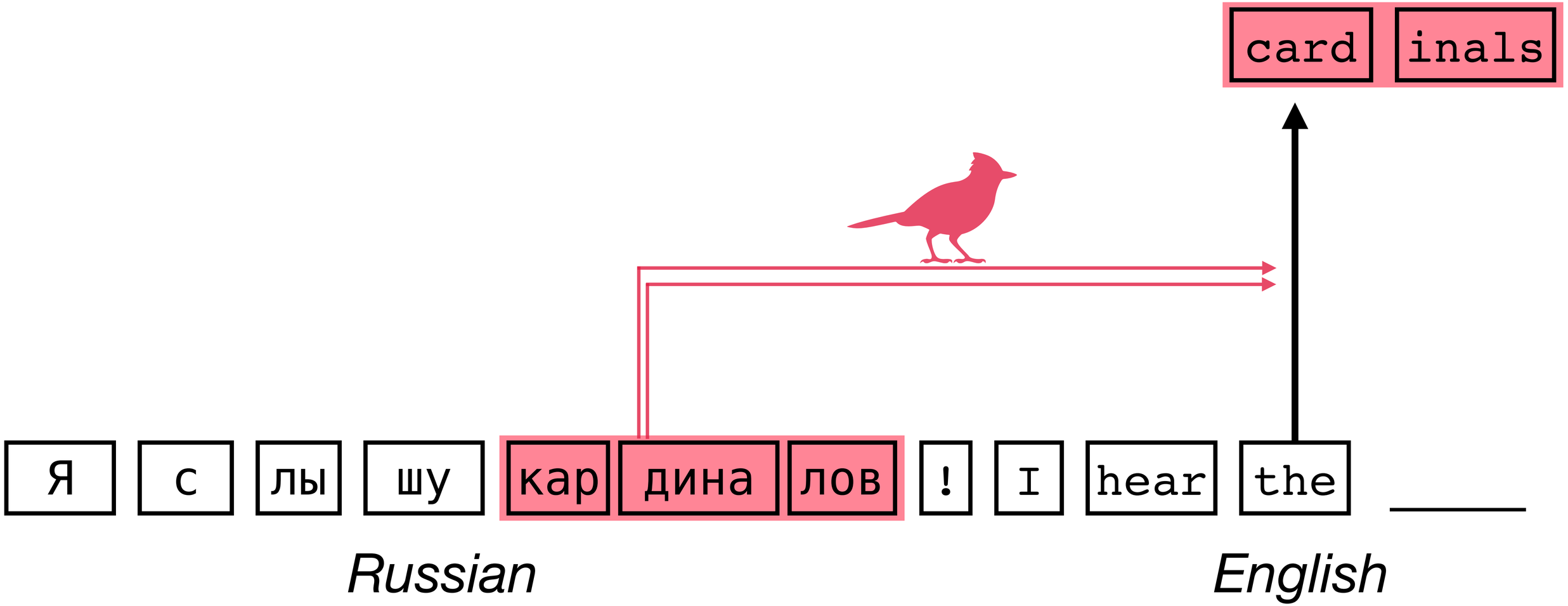
We hypothesize that these heads are copying the **meaningful** word “cardinals”.



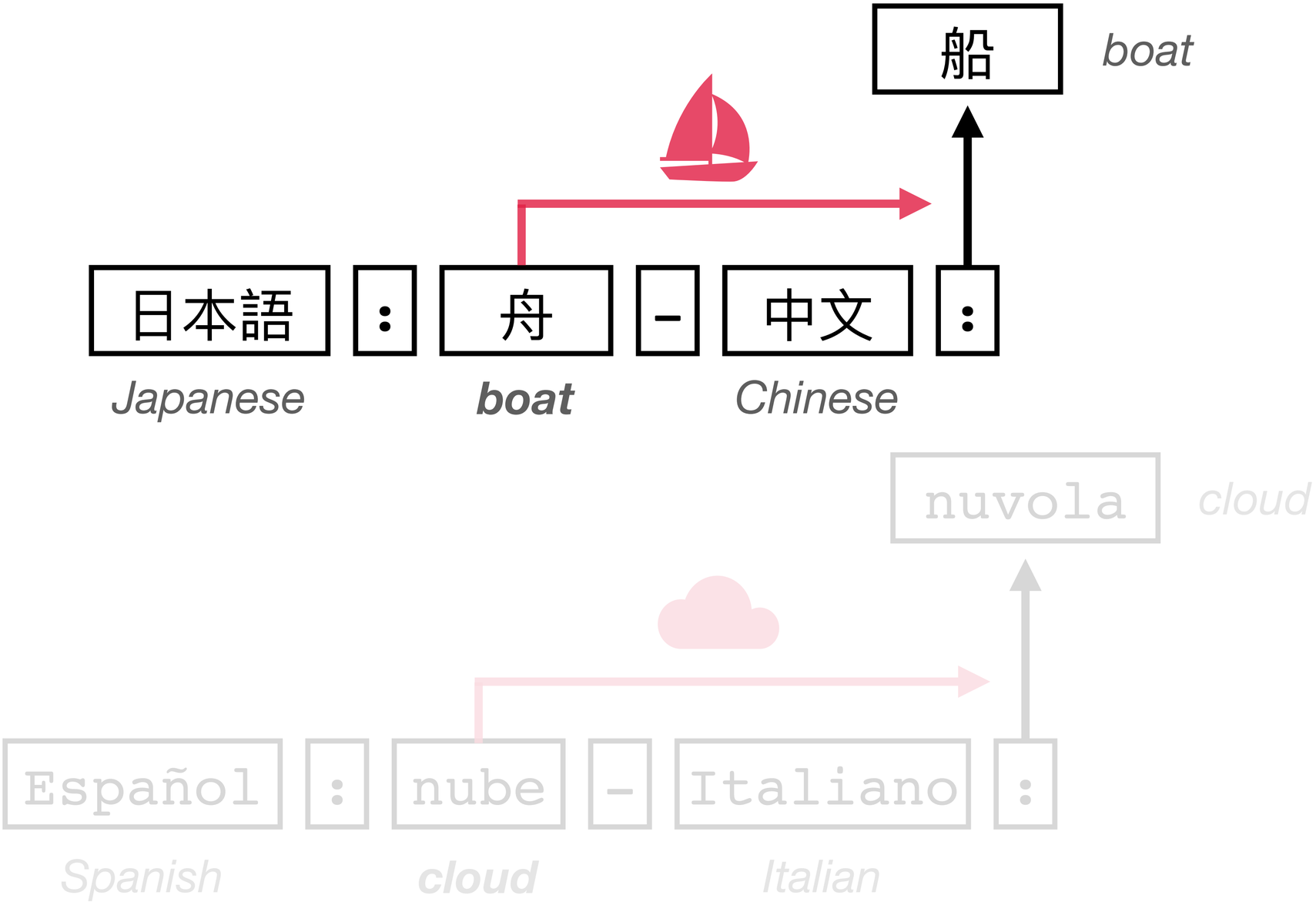
If concept induction heads copy word **meanings**, they should function the same regardless of how a word is written.



If concept induction heads copy word **meanings**, they should function the same regardless of how a word is written.

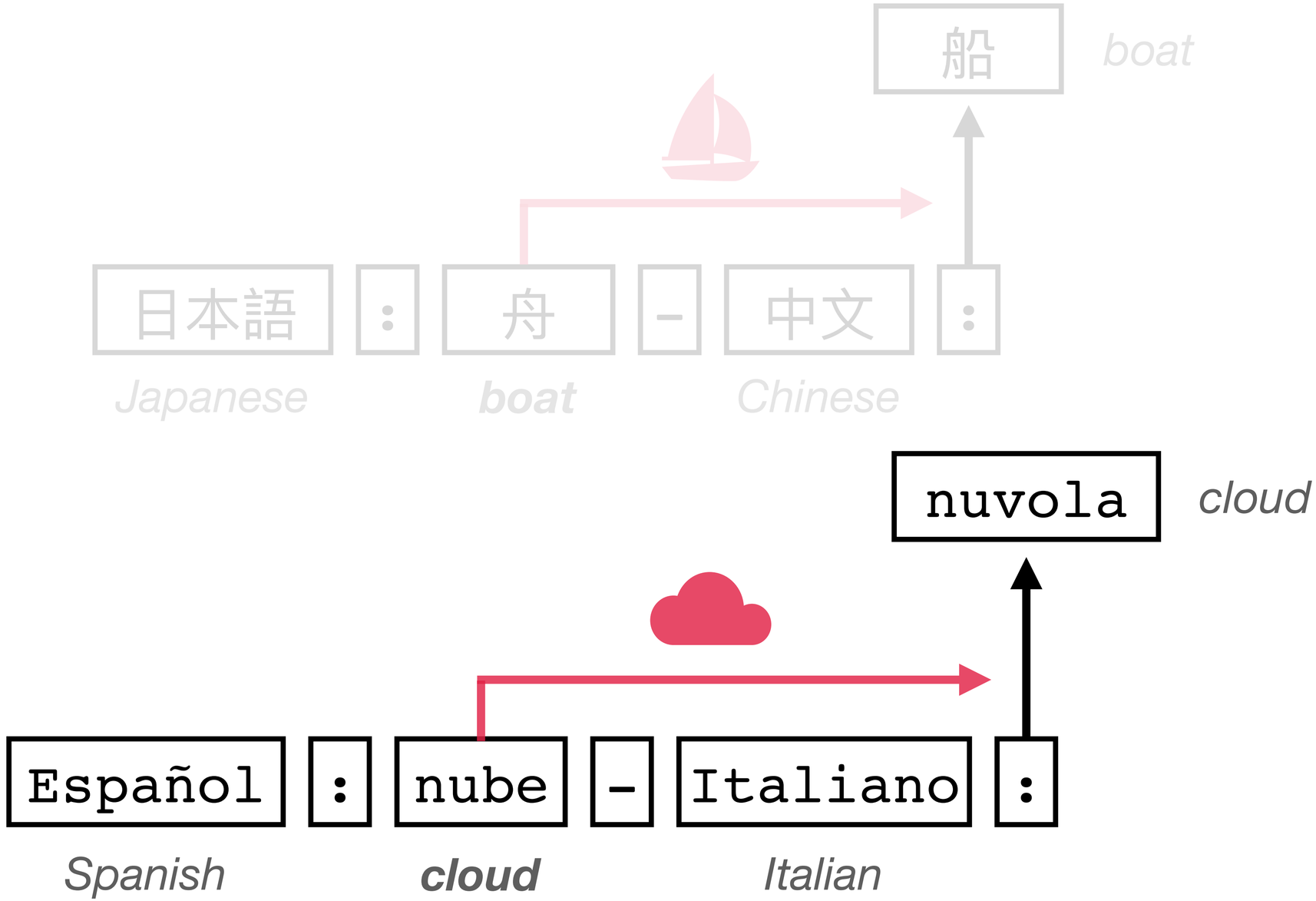


We find this is true! They are used to translate words.



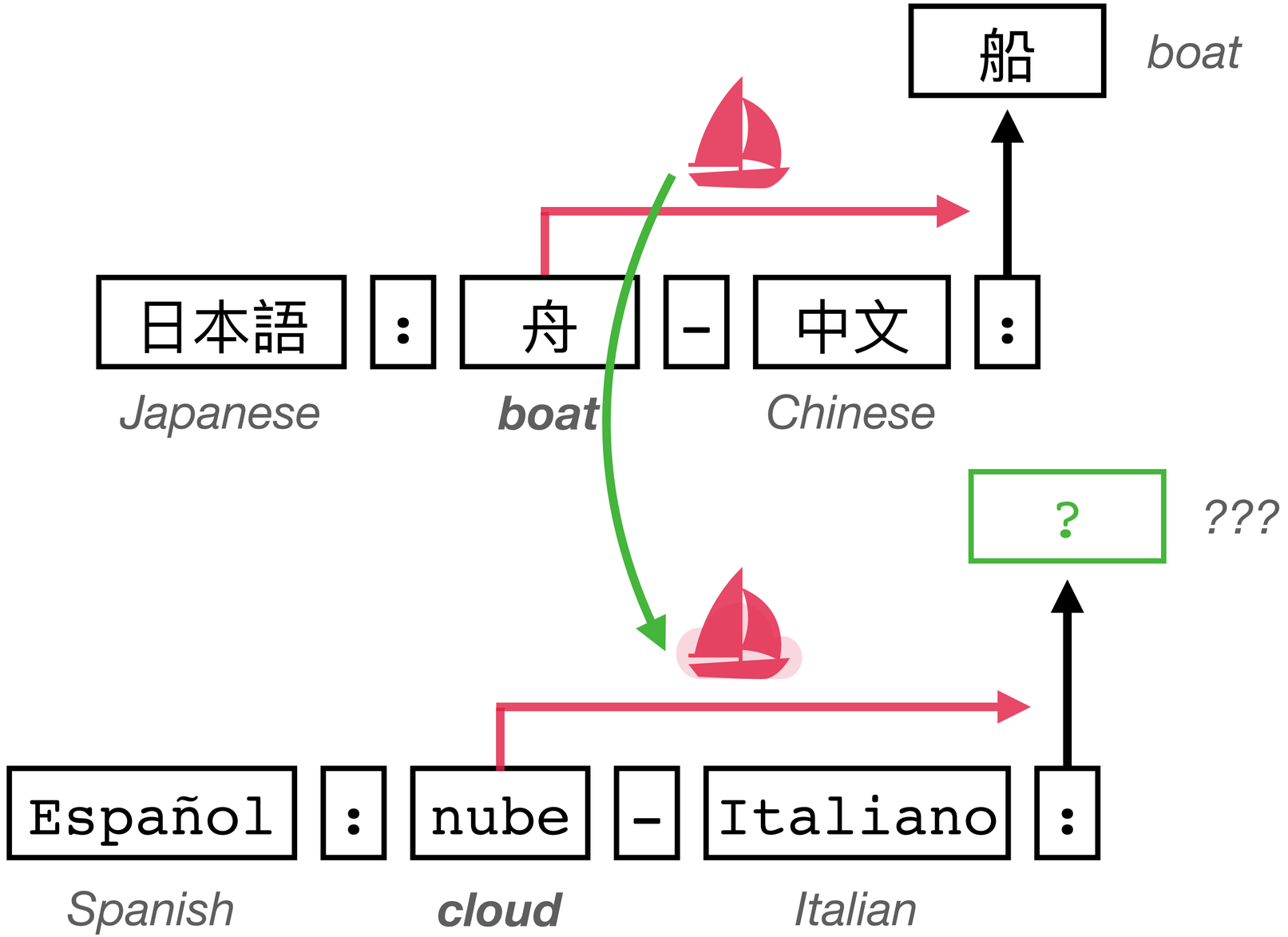
Setup from Clément Dumas et al. (2024) - ICML 2024 Workshop on Mechanistic Interpretability

We find this is true! They are used to translate words.



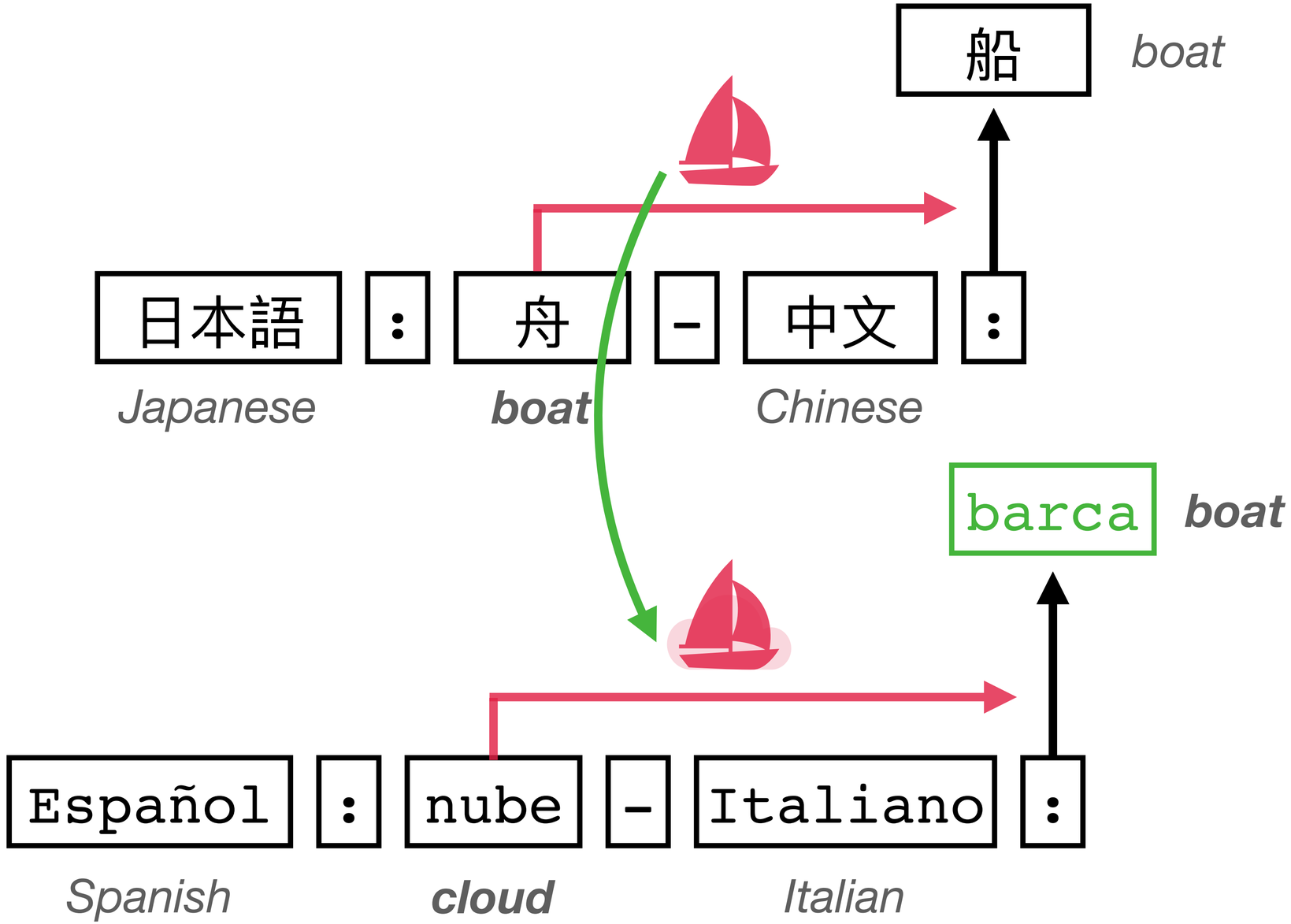
Setup from Clément Dumas et al. (2024) - ICML 2024 Workshop on Mechanistic Interpretability

What happens if we patch these heads into a new context?



Setup from Clément Dumas et al. (2024) - ICML 2024 Workshop on Mechanistic Interpretability

It causes the model to output “boat” in Italian!



Setup from Clément Dumas et al. (2024) - ICML 2024 Workshop on Mechanistic Interpretability