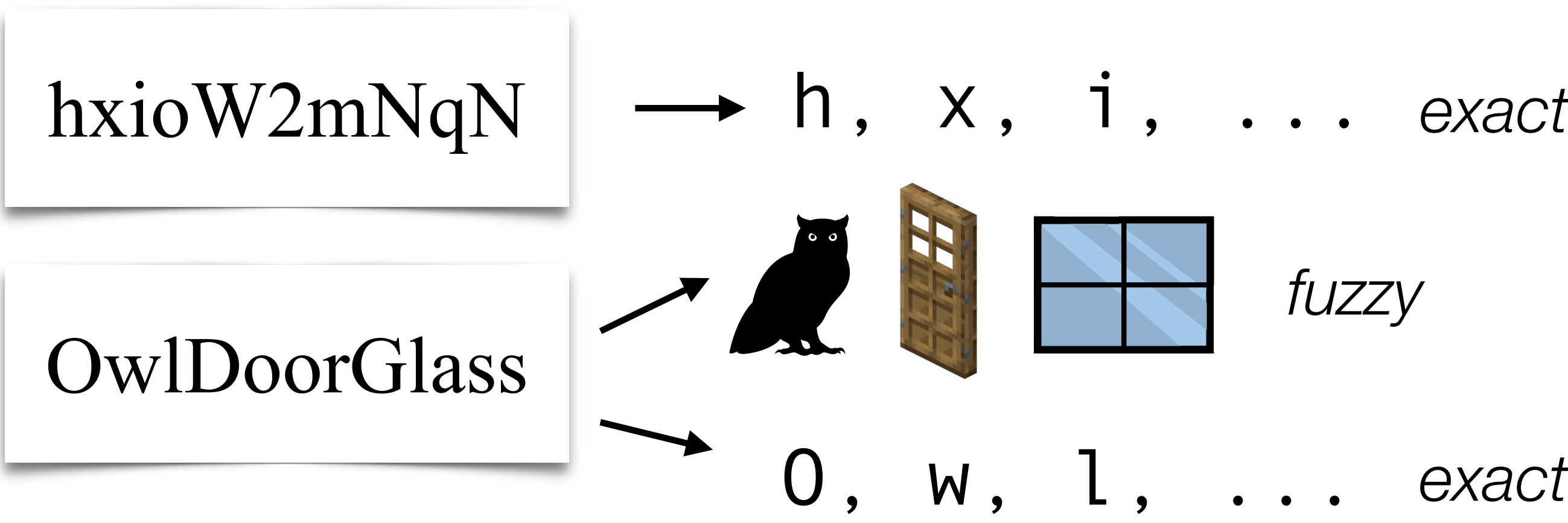
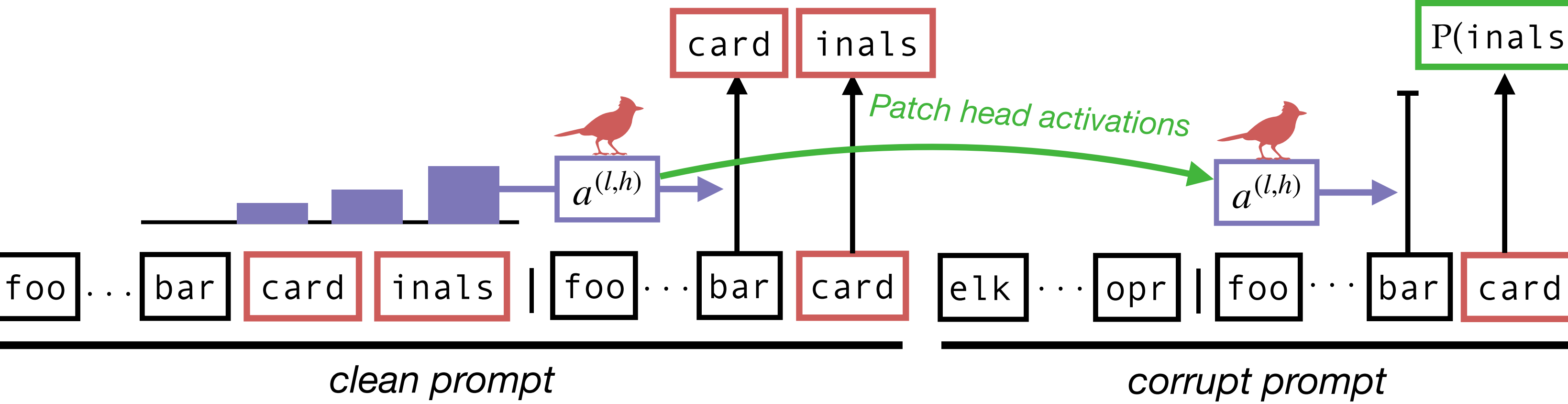


“Fuzzy” copying: we don’t look at every single letter when copying a Wi-Fi password.

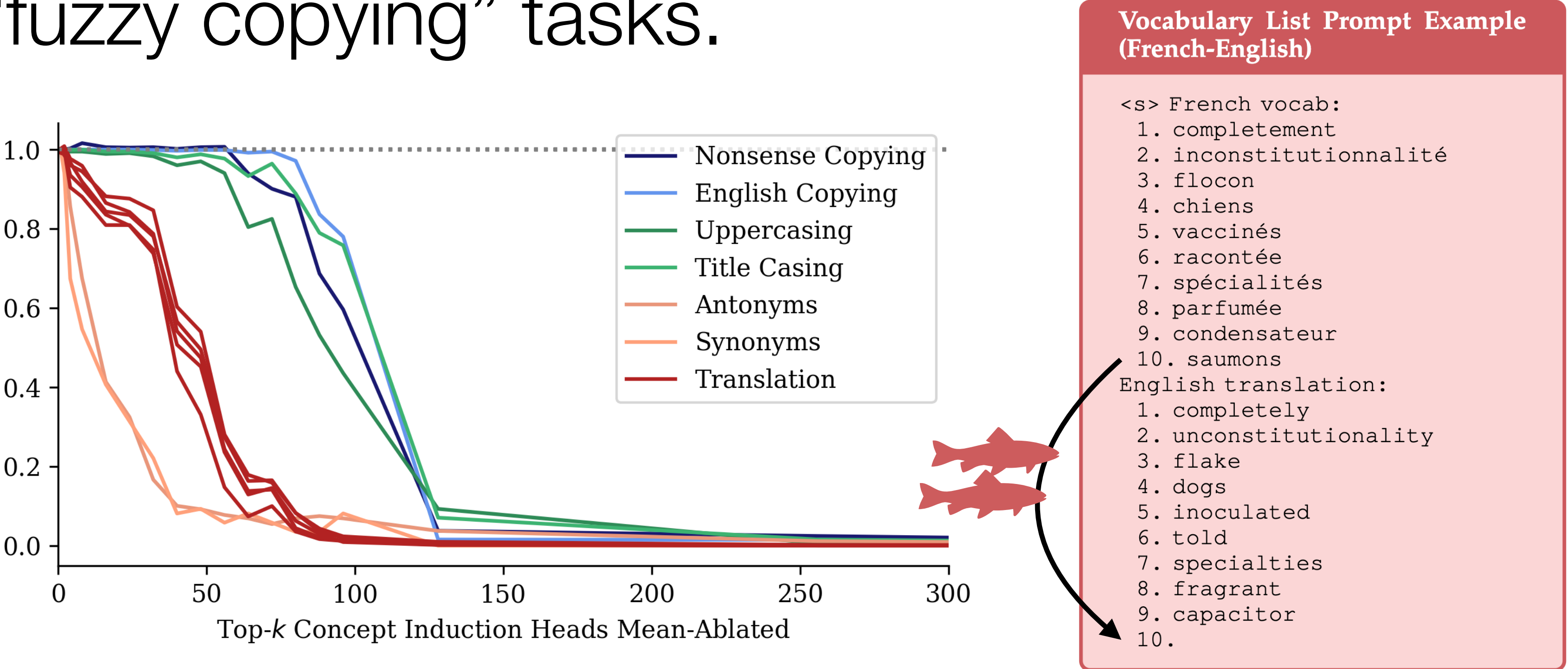


Do LLMs do this? We look for heads that copy multiple tokens at a time.

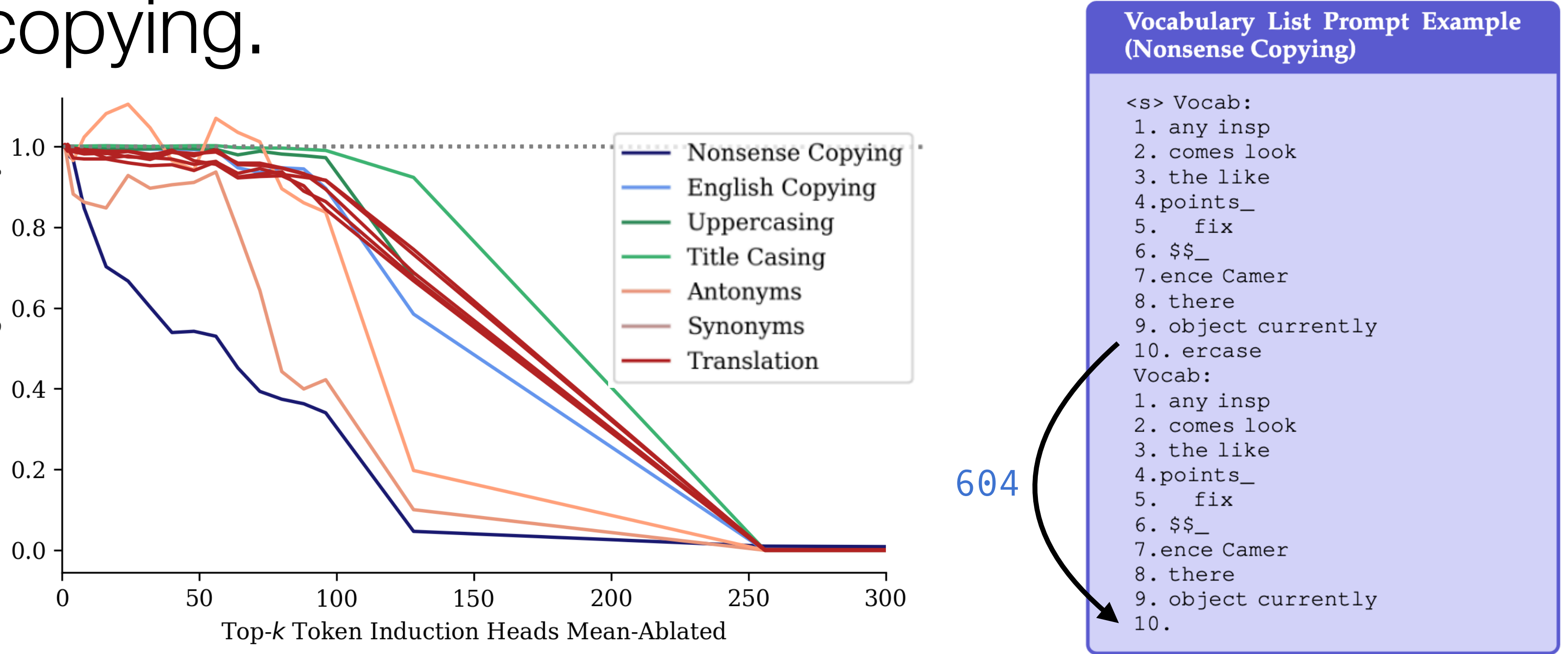


Patch each head into a context without the word “cardinals,” and see if P(inals | card) increases.

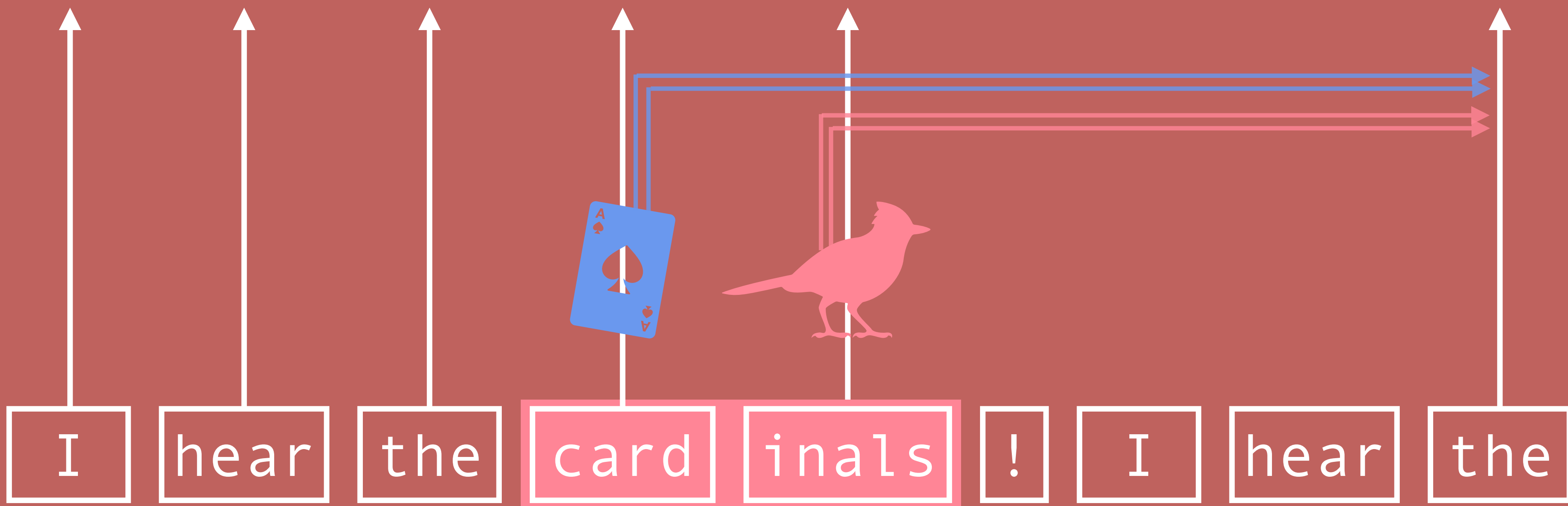
Ablating these **concept** heads damages “fuzzy copying” tasks.



Ablating **token** heads damages verbatim copying.




LLMs copy two ways:



by *token* and by *concept*.

The Dual-Route Model of Induction

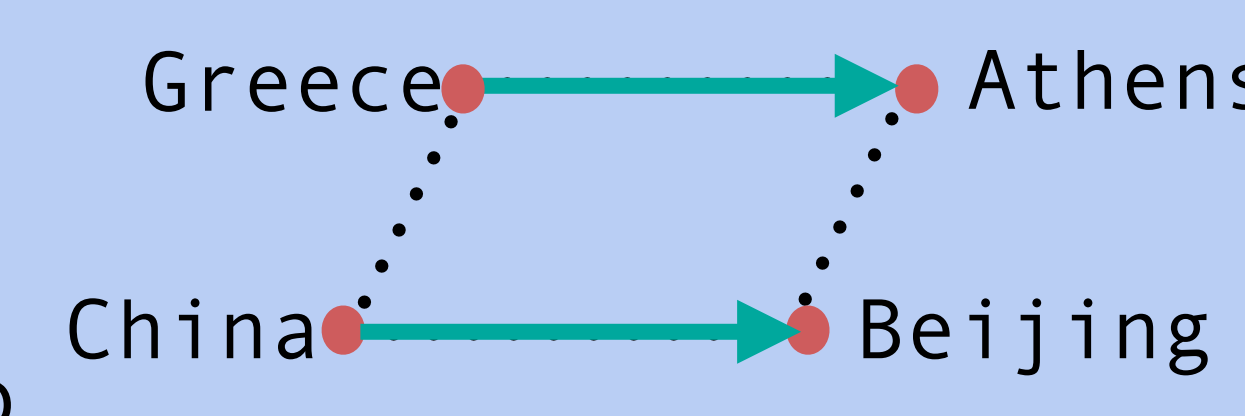
Sheridan Feucht, Eric Todd, Byron Wallace, David Bau
Northeastern University



Bonus: word2vec arithmetic with concept/token vectors!

<https://arithmetic.baulab.info>

<https://dualroute.baulab.info>



Meaningful words can be copied **both** ways. But without token heads, models start to paraphrase.

Vocabulary List Prompt Example (English)

```
<s> English vocab:
1. live
2. begin
3. stumble
4. good
5. ostracize
6. important
7. coin
8. colored
9. manner
10. recover
English vocab:
1. live
2. begin
3. stumble
4. good
5. ostracize
6. important
7. coin
8. colored
9. manner
10.
```

Llama-2-7b: Copying Text

I have reread, not without pleasure, my comments to his lines, and in many cases have caught myself borrowing a kind of opalescent light from my poet’s firey orb.

Top-32 Token Heads Ablated

I have reread my comments on his lines, and I have caught myself many times borrowing from his firey orb a kind of opalescent light.

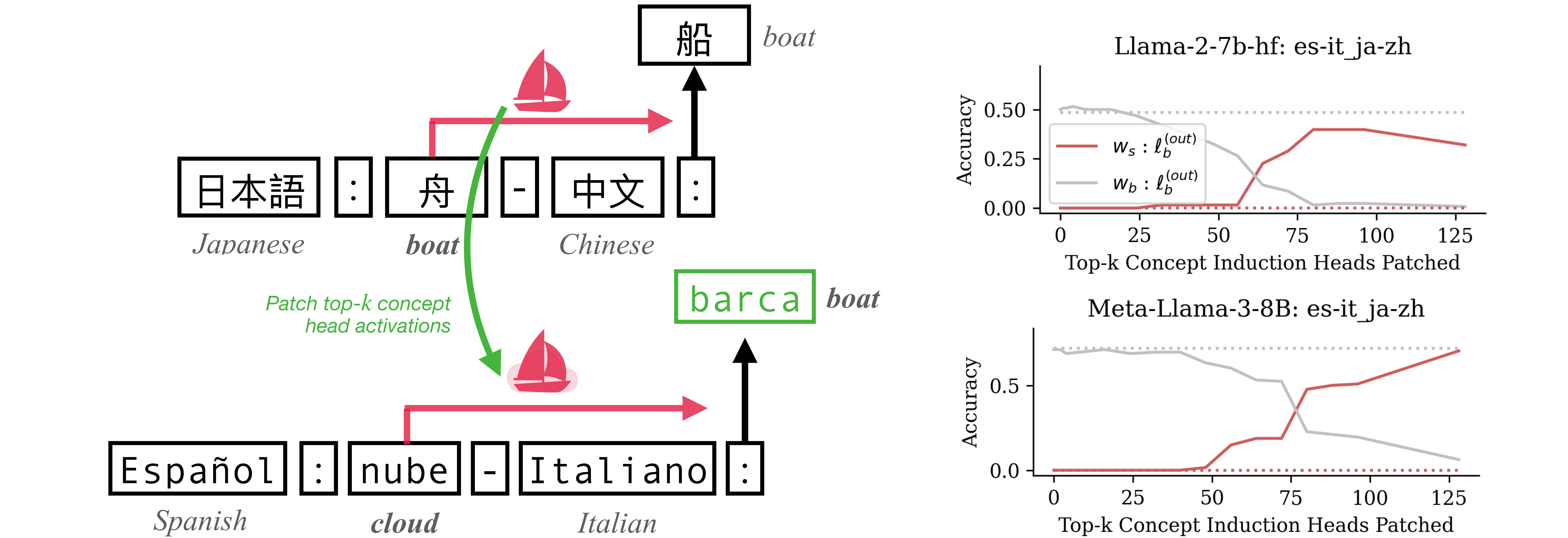
Llama-3-8b: Copying Code

```
foo = []
for i in range(len(bar)):
    if i % 2 == 0:
        foo.append(bar[i])
```

Top-32 Token Heads Ablated

```
foo = [bar[i] for i in
range(len(bar)) if i % 2 == 0]
```

Concept heads copy “language-independent” representations—they copy the meaning of the word, not the literal tokens in that word!



We can use concept head weights to reveal word semantics within arbitrary hidden states.

We sum the top-*k* concept OV matrices (kind of like forcing all heads to attend to this hidden state.)

$$s_l = s_{l-1} + \sum_{(l,h)} s_{l-1} V_{(l,h)} O_{(l,h)}$$

