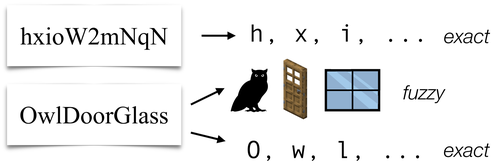
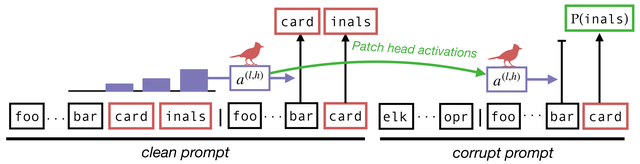
"Fuzzy" copying: we don't look at every single letter when copying a Wi-Fi password.

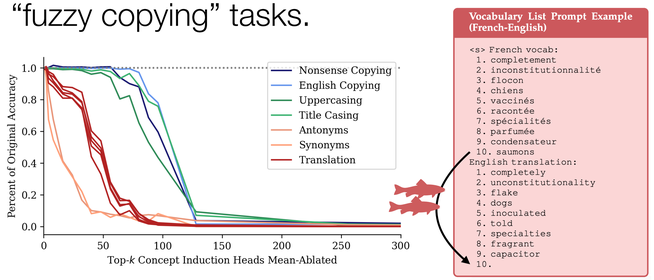


Do LLMs do this? We look for heads that copy multiple tokens at a time.

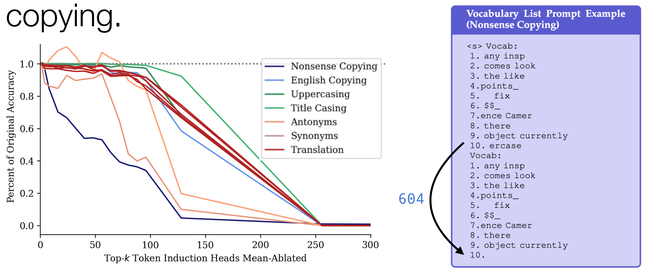


Patch each head into a context without the word "cardinals," and see if P(`inals|card`) increases.

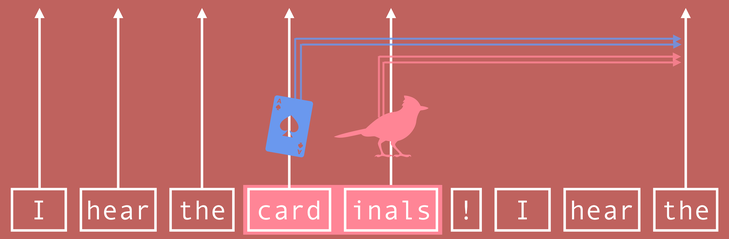
Ablating these **concept** heads damages "fuzzy copying" tasks.



Ablating **token** heads damages verbatim copying.



LLMs copy two ways:

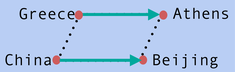


*by **token** and by **concept.***

The Dual-Route Model of Induction

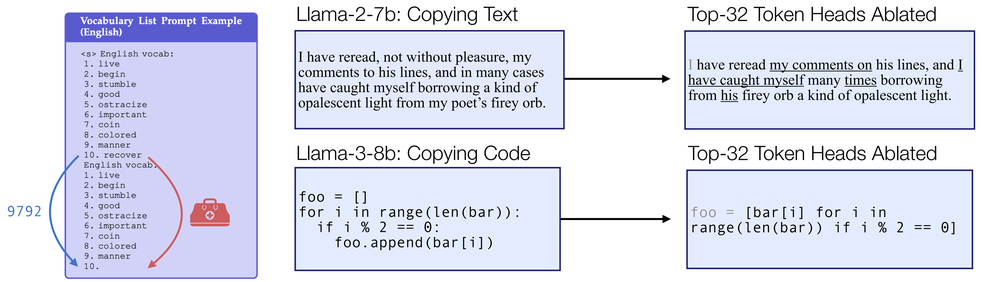
Sheridan Feucht, Eric Todd, Byron Wallace, David Bau
Northeastern University

Bonus: word2vec arithmetic with concept/token vectors!
https://arithmetic.baulab.info

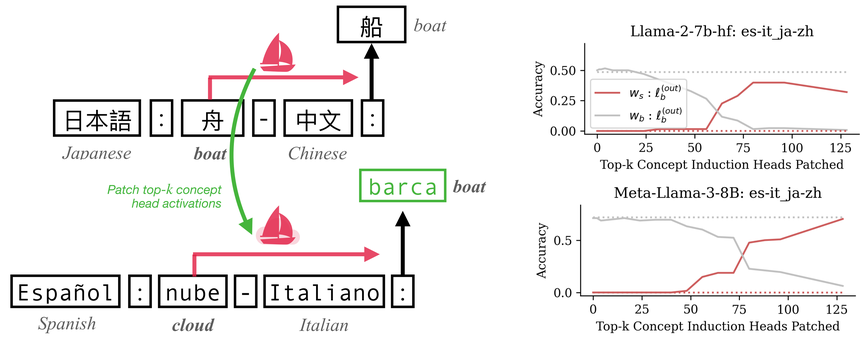


https://dualroute.baulab.info

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



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.)

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

Instead of doing raw logit lens,

`lm_head(` s `)`

We first apply "concept lens,"

`lm_head(` s $\sum VO$ `)`

which reveals semantic information in these hidden states.