Probing and **causal interventions** yield different accounts of **LMs' processing**

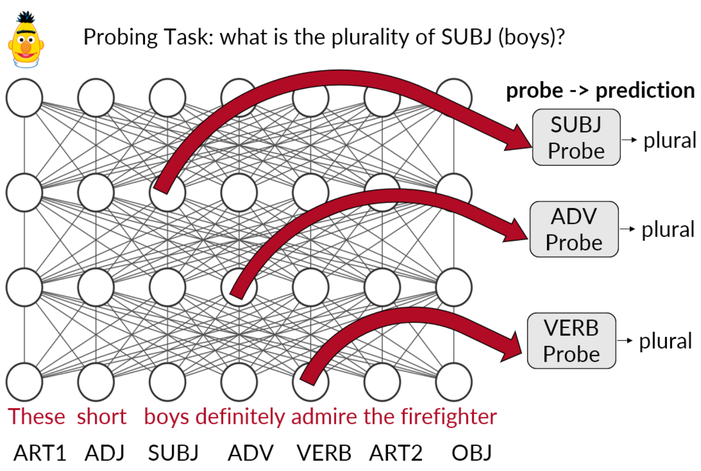
The Functional Relevance of Probed Information: A Case Study

Michael Hanna, Roberto Zamparelli, David Mareček
m.w.hanna@uva.nl

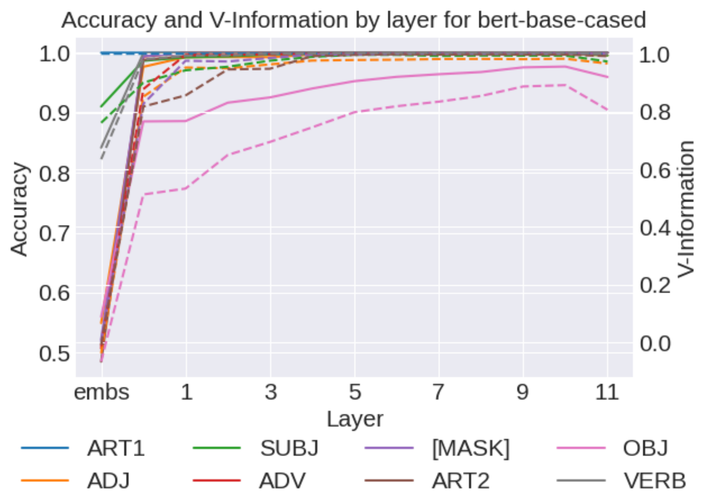
University of Amsterdam
Institute for Logic, Language, and Computation
Amsterdam, Netherlands

Introduction

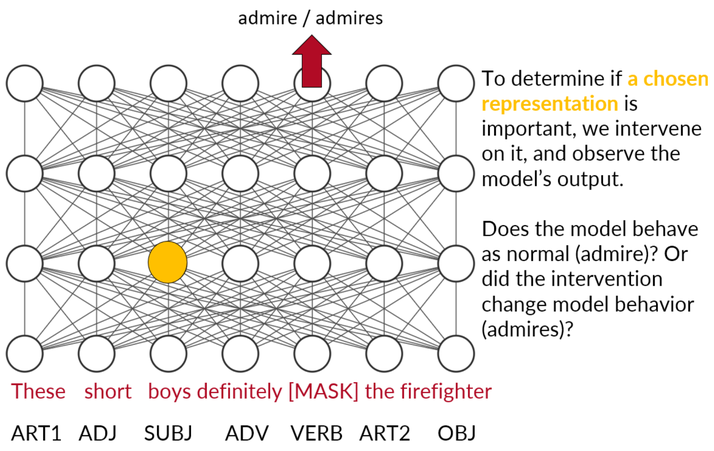
Transformer language models (LMs) mix information across different token positions. This can be detected via probing.



SUBJ plurality information is encoded in **all** tokens [1]. This is **linguistically implausible**.

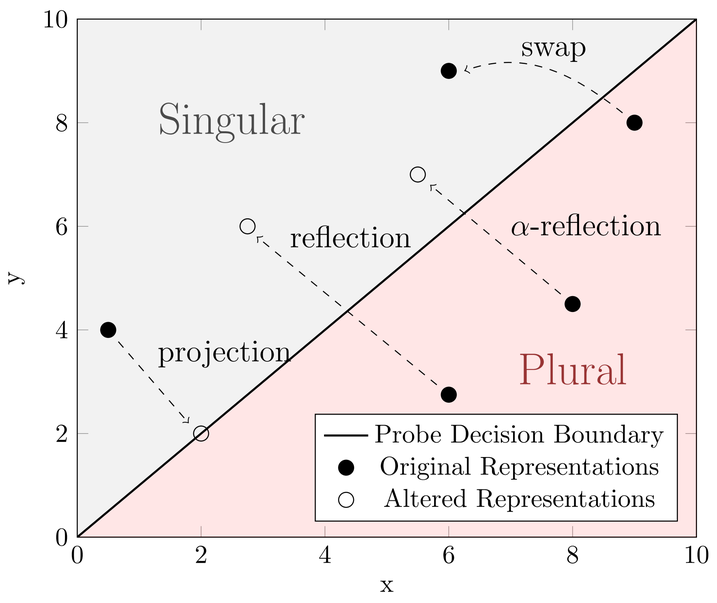


But is this encoded info actually used? We investigate using causal interventions:



Causal Interventions

Using our linear probes, **we change SUBJ plurality info** in word representations:

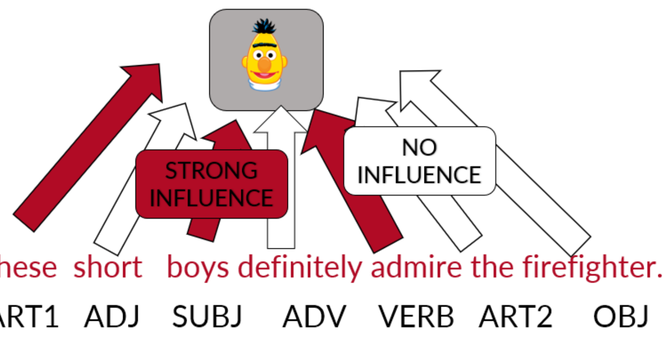


- **Projection**: Removes plurality information
- **Reflection**: Flips plurality information
- **α-reflection**: (Barely) flips plurality [2]
- **Swap**: Flips plurality by replacing the word representation with a representation of the same word, with opposite plurality [3]

Successful interventions increase disagreement between the subject noun and the verb (conjugation) predicted by the model.

Results

To what extent does each word's plurality information influence LM behavior?

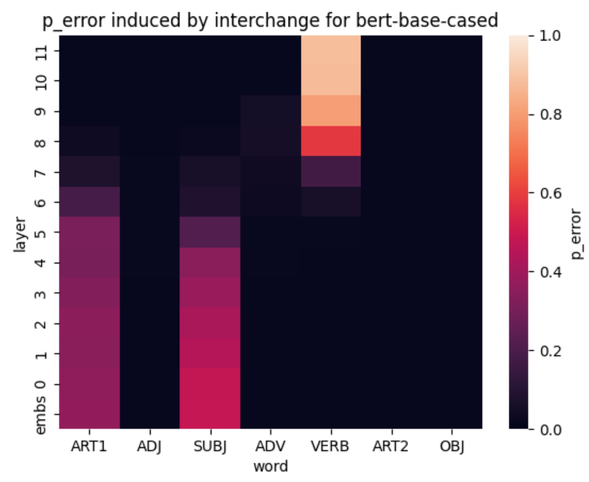


- Swapping produces large effects, while reflection produces slightly smaller ones.
- Projection and α-reflection are ineffective.

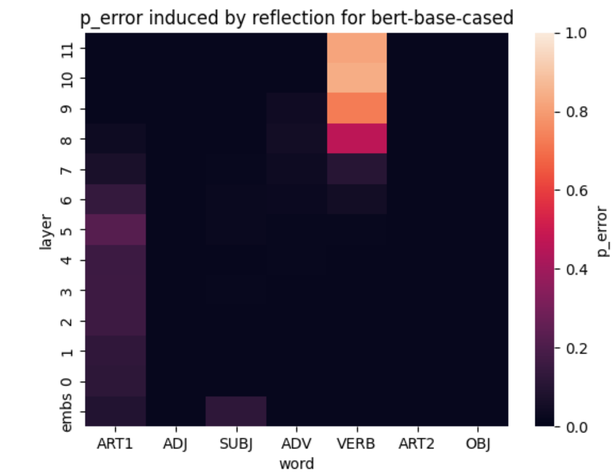
Data and Details

We train linear probes on 4k sentences like "[This/these] [adjective] [subject] [adverb] [verb] the [object]." to predict subject number. We intervene on LMs that predict the masked verb of such sentences. To evaluate, we calculate the probability of verb forms that disagree with the subject, post-intervention.

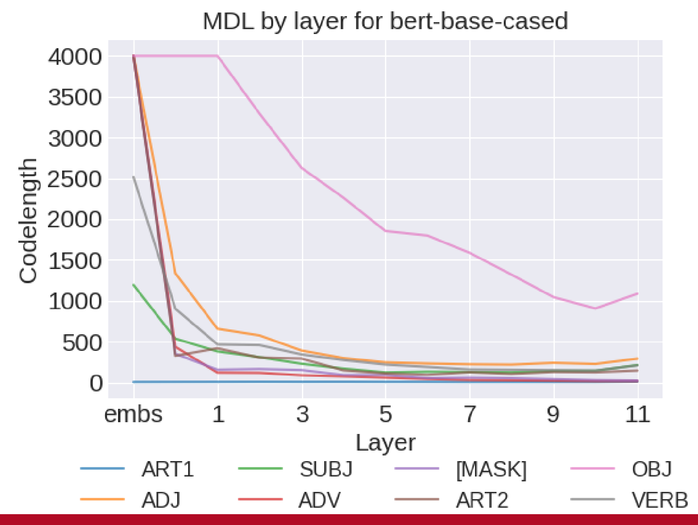
Interventions on the subject and words that agree with it succeed, indicating that LMs use information stored there.



Reflection interventions are less effective, indicating that although information stored at the subject position is used, it may not be used as found by the probe.



We try to use $\mathcal{V}$-information (left) and minimum description length probing to detect functional relevance, but these fail.



References

[1]: Josef Klafka and Allyson Ettinger. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. July 2020. Association for Computational Linguistics.

[2]: Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. November 2021. CoNLL.

[3]: Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. November 2020. BlackBoxNLP

Master's thesis project at the University of Trento and Charles University, supervised by Profs. Roberto Zamparelli and David Mareček. Poster template by Mike Morrison, Rafael Bailo, Tom Kocmi.