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

Probing Interventions on Nominal Plurality Representations in LLMs

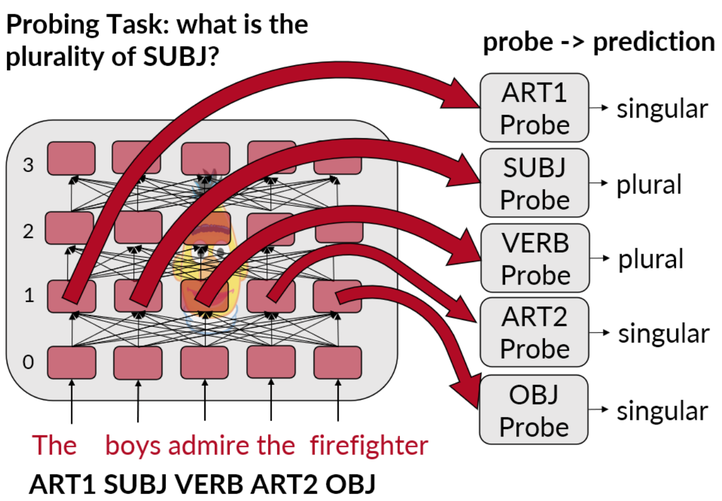
Michael Hanna

michaelwesley.hanna@studenti.unitn.it

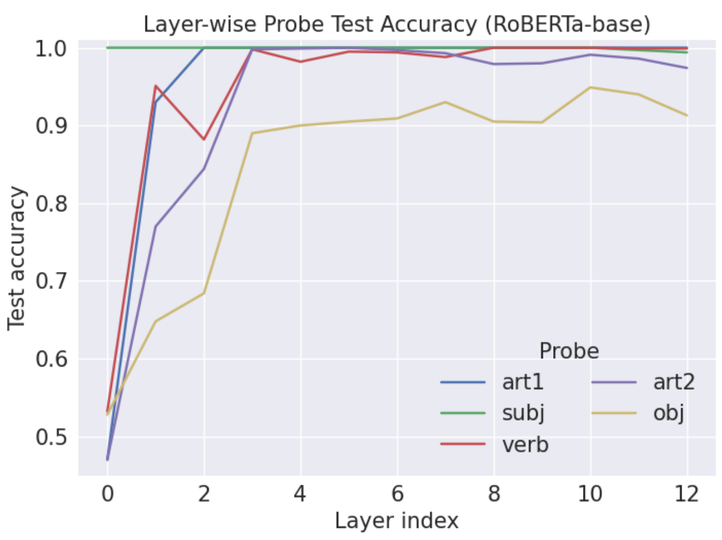
University of Trento, Department of Cognitive Science and Pschology
Center for Mind/Brain Sciences
Rovereto, Italy

Introduction

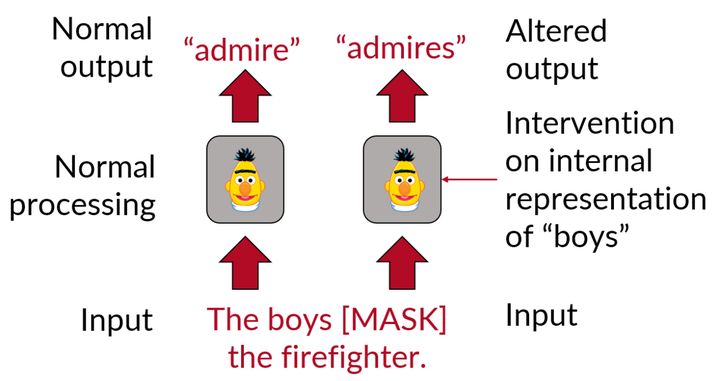
Large language models' (LLMs) transformers enable **cross-token information mixing**. This can be detected via probing.



SUBJ plurality information is encoded in **all** tokens—this is **linguistically implausible**.

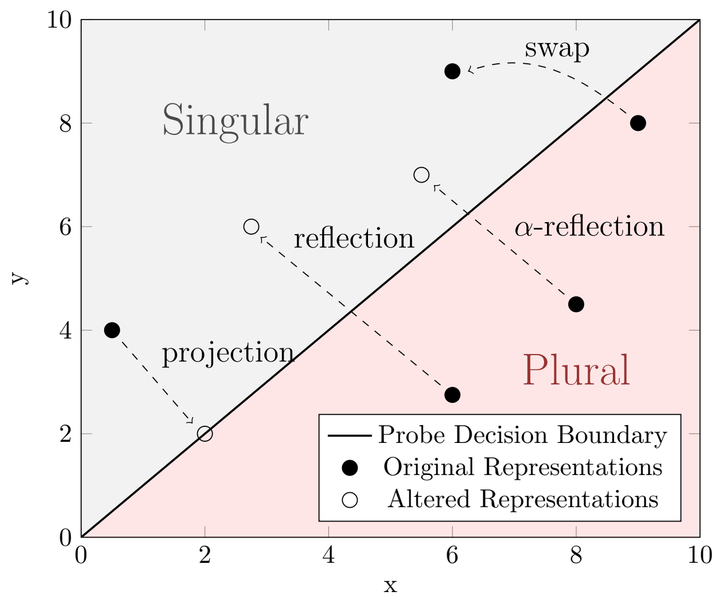


But, is this encoded info actually used? We investigate using causal interventions:



Causal Interventions

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

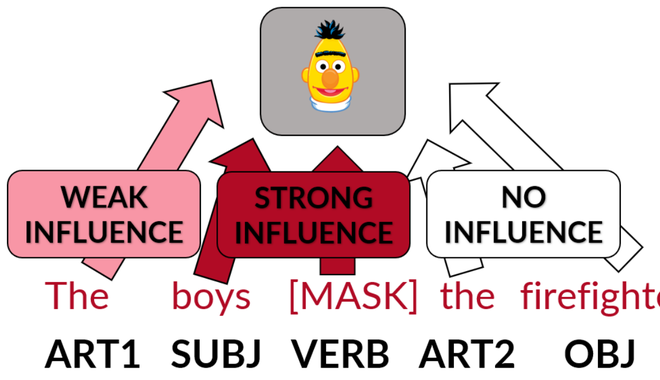


- **Projection**: Removes plurality information
- **Reflection**: Flips plurality information
- **α-reflection**: (Barely) flips plurality [2]
- **Swap**: Flips plurality by replacing the word representation with a representation of the same word, with opposite plurality [3]

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 LLM behavior?

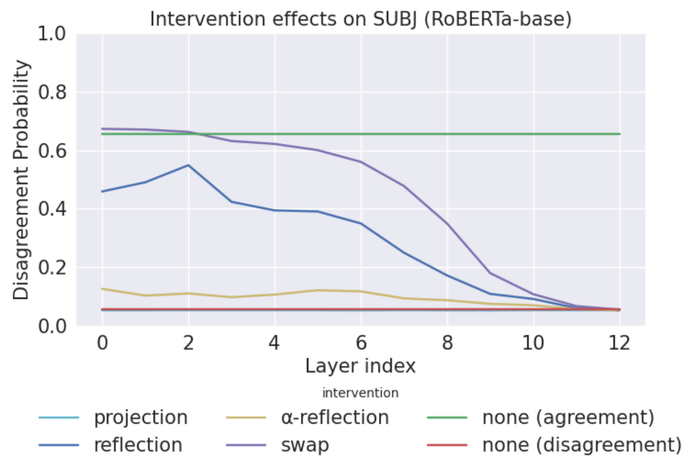


- Swapping produces large effects, while reflection produces slightly smaller ones.
- Projection and α-reflection are ineffective.

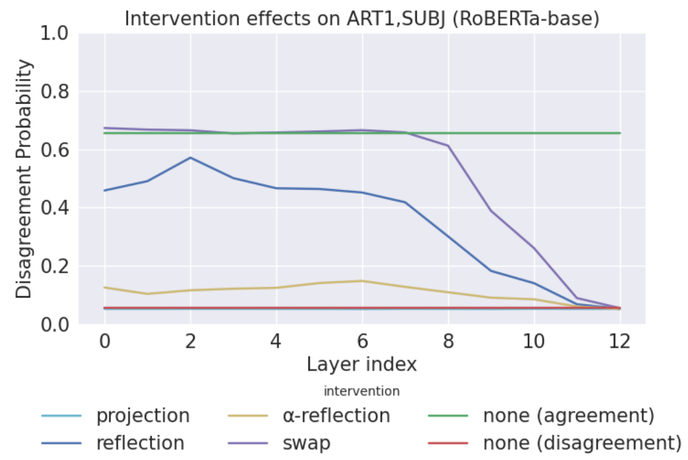
Data and Details

We train linear probes on 4000 sentences of the form "The [subject] [verb] the [object]." to predict the subject's plurality. We then intervene on LLMs predicting the masked verb of such sentences. We evaluate by calculating the probability mass assigned to verb forms that disagree with the subject, pre- and post- intervention.

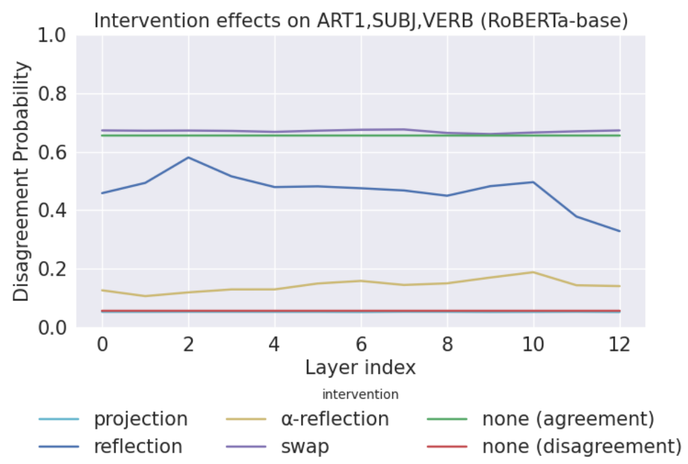
Interventions on the subject yield significant effects for swapping and reflection, but little for α-reflection and none for projection. Effects weaken in later layers.



Intervening on ART1 as well increases later-layer effects slightly.



Adding VERB makes swap disagreement match the original agreement. Reflection effects in later layers increase too.



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.