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

Probing Interventions on Nominal Plurality Representations in LLMs

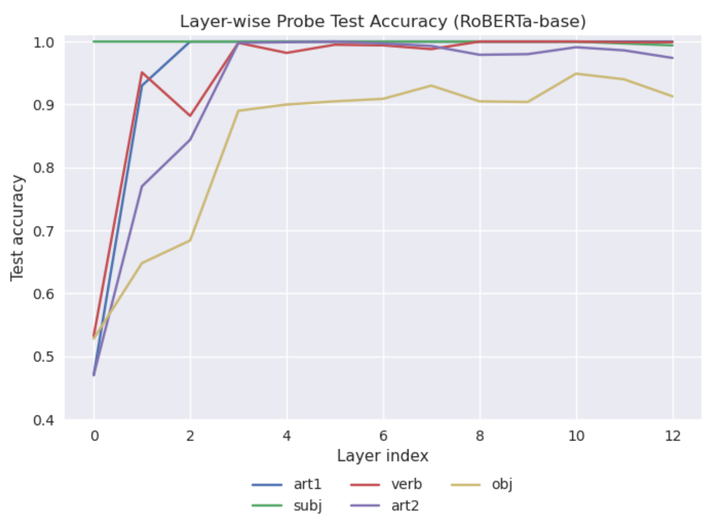
Michael Hanna
michaelwesley.hanna@studenti.unitn.it

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

Introduction

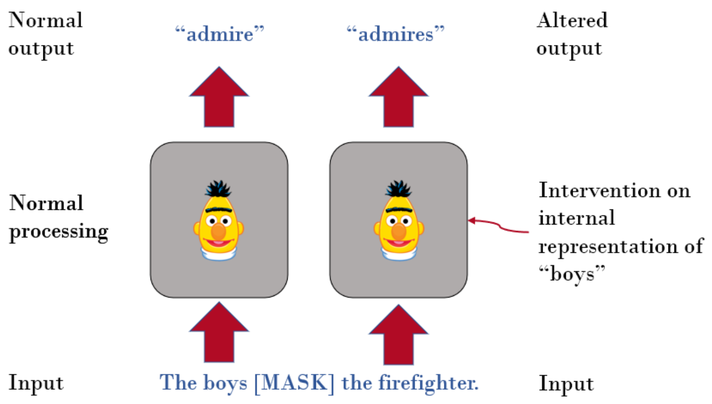
Large Language Models (LLMs) use transformers, creating token representations by attending to all tokens in the sentence. This can cause undesirable information mixing.

Using LLM representations of simple 5-word sentences such as "The lawyers like the judge," we train 5 linear probes. Each predicts the plurality of a sentence's subject noun from representations at one word position. They can extract plurality info from all positions, and almost every layer. [1]

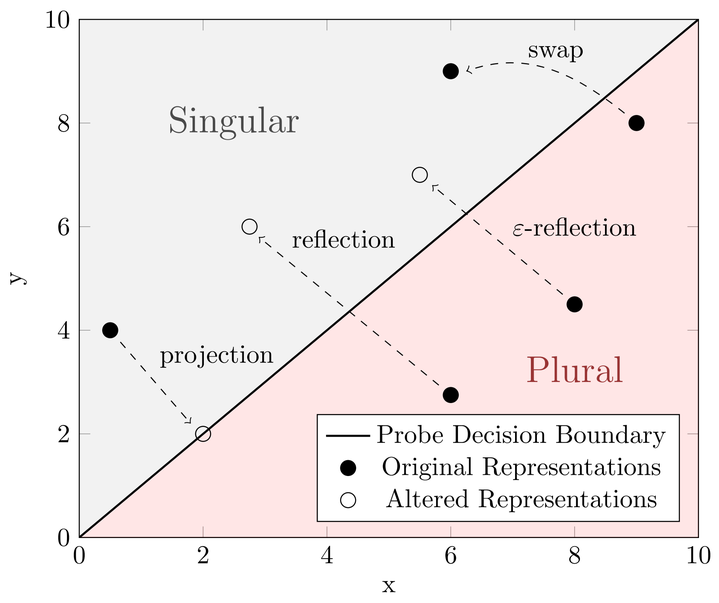


Subject noun plurality should not be encoded in all words in the sentence. But is this stored information actually used? We investigate with causal interventions.

Causal Interventions



Using the linear probes trained earlier, we intervene to change subject plurality information in each word representation:



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

We evaluate based on model behavior. Successful interventions will increase disagreement between the verb (conjugation) predicted by the model and the subject noun.

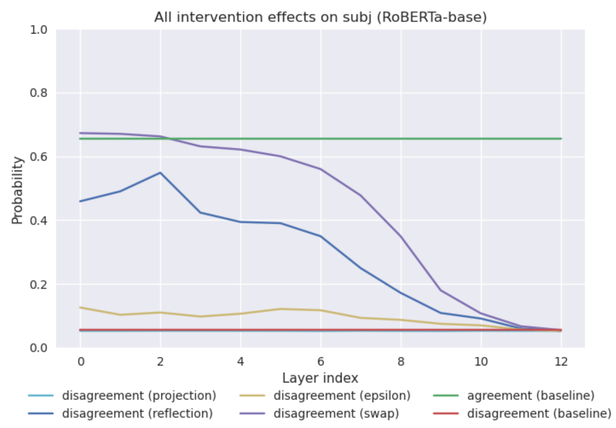
Results

- Plurality info in the subject is most used; the verb contributes in later layers.
- Plurality info encoded in the subject's article is only slightly used.
- Plurality info encoded in the object and its article is not used at all.
- Swapping produces large effects, while reflection produces slightly smaller ones.
- Projection and ε-reflection are ineffective.
- This replicates across different LLMs.

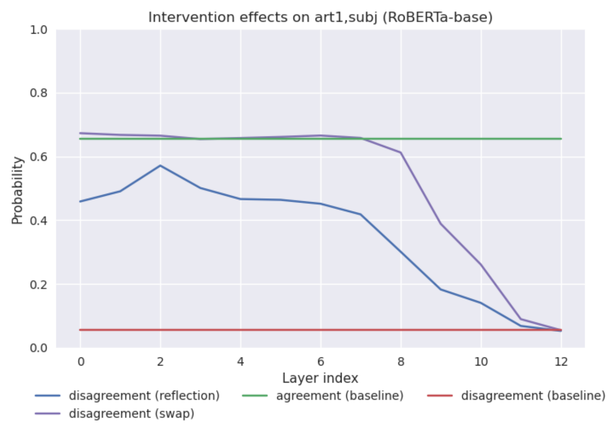
Data and Details

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.

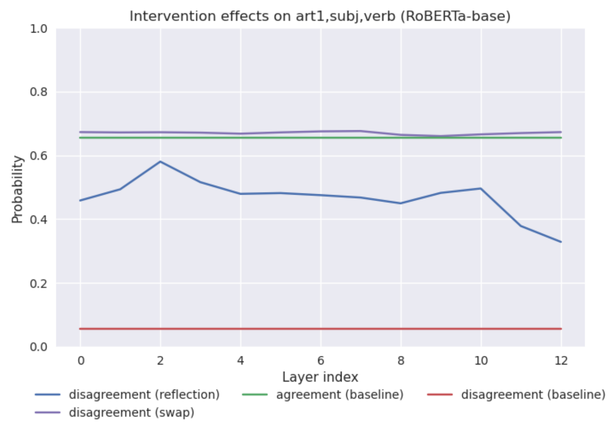
Interventions on the subject yield significant effects for swapping and reflection, but little for ε-reflection and none for projection. Effects weaken in later layers.



Intervening on the article as well increases later-layer effects slightly.



Also intervening on the verb yields swap effects of the same magnitude as the original agreement. Reflection effects in later layers increase, but remain below swap.



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.