Discovering Latent Knowledge in Language Models Without Supervision - extensions and testing

Agatha Duzan, Matthieu David, Jonathan Claybrough

AI Testing Hackathon Report

Date: 18th December, 2022

Abstract

Based on the paper "Discovering Latent Knowledge in Language Models without supervision" this project discusses how well the proposed method applies to the concept of ambiguity.

To do that, we tested the Contrast Consistent Search method on a dataset which contained both clear cut (0-1) and ambiguous (0,5) examples : we chose the ETHICS-commonsense dataset. The global conclusion is that the CCS approach seems to generalize well in ambiguous situations, and could potentially be used to determine a model's latent knowledge about other concepts.

Discovering Latent Knowledge in Language Models Without Supervision - extensions and testing.

Our work seeks to improve upon the paper "Discovering Latent Knowledge in Language Models without supervision". This paper introduces a novel method to determine a language model's concept of "truth", which enables evaluating a prompt's truthfulness simply by looking at the model's representation of that prompt (in the last layer) and going through a probe. This kind of advance in interpretability is notable for presenting a general approach of discovering latent knowledge without supervision. Our project aimed to test the limits of this approach, as well as extend it.

In the first place, we tried to replicate the results of the paper on our own : we found that on smaller language models (deBerta and GPT-J), the distinction between 'probably true' and 'probably false' examples was less clear than in the paper (figures 1, 2 and 3). This may be due to the model used by the author. C Burns *et al.* [1] used a version of T5 fine tuned for question/answers format).

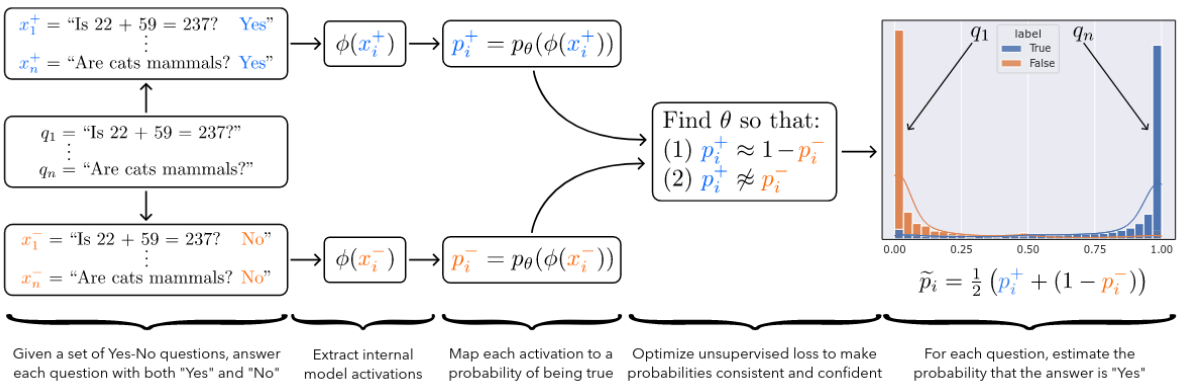

Figure1: Figure from the article of C. Burns et al. On the right hand side, one can observe a histogram of the "Yes" probabilities, ***$\tilde{p}i = 0.5 \cdot (p+i + (1 - p-i))$****, learned by the method for the COPA dataset (Roemmele et al., 2011) with the UnifiedQA model (Khashabi et al., 2020)*

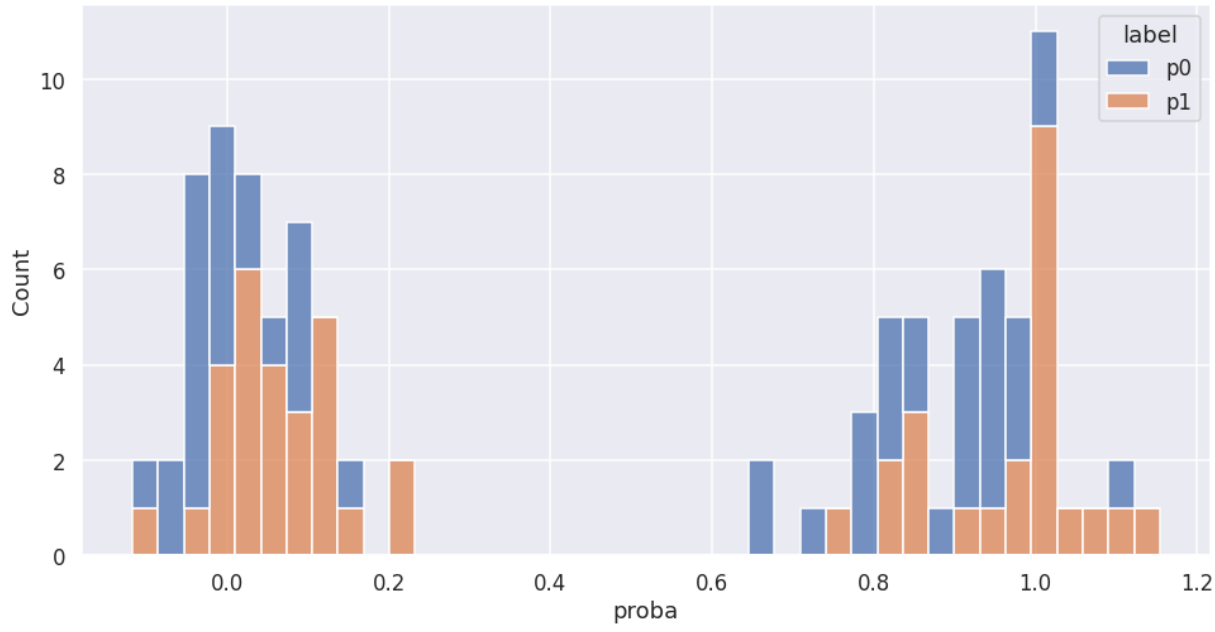

Figure2: CSS results for last layer activation of Deberta model (pretrained model from transformers python library) for the test set of the dataset Amazon polarity. Results can be found in the [github repository](#).

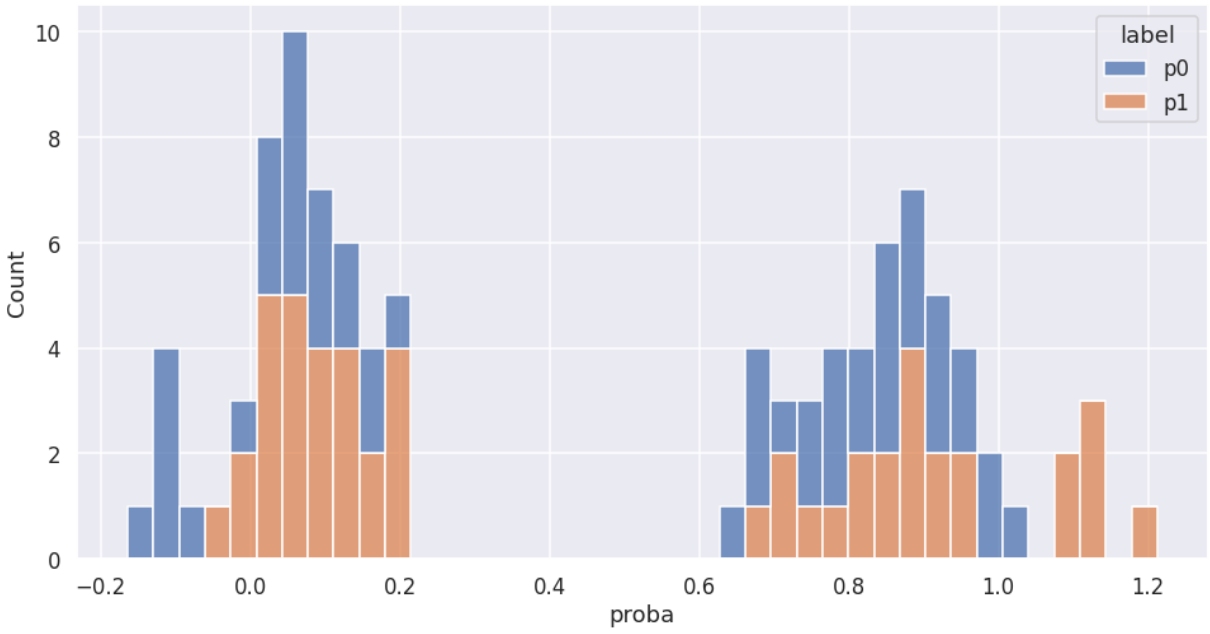

Figure3: CSS results for last layer activation of GPT-J model (pretrained model from transformers python library) for the test set of the dataset Amazon polarity. Results can be found in the [github repository](#).

Secondly, we applied CCS to the ETHICS-commonsense dataset : the training was done on exclusively clear-cut examples, and we tested the accuracy on the ambiguous testing examples.

Earlier, on the clear-cut examples, the probability plot formed two bins : one on the far left corresponding to ‘probably false’ examples, and one on the far right corresponding to ‘probably true’ examples.

The hypothesis we wanted to test was whether the middle zone of the probability graph would correspond to ‘ambiguity’. We observed the following phenomenon on the probability graph made on the ambiguous examples : both distributions of ‘positive’ and ‘negative’ probabilities were approximately centred in the middle, and the underlying bell curve was much flatter than in the earlier unambiguous case.

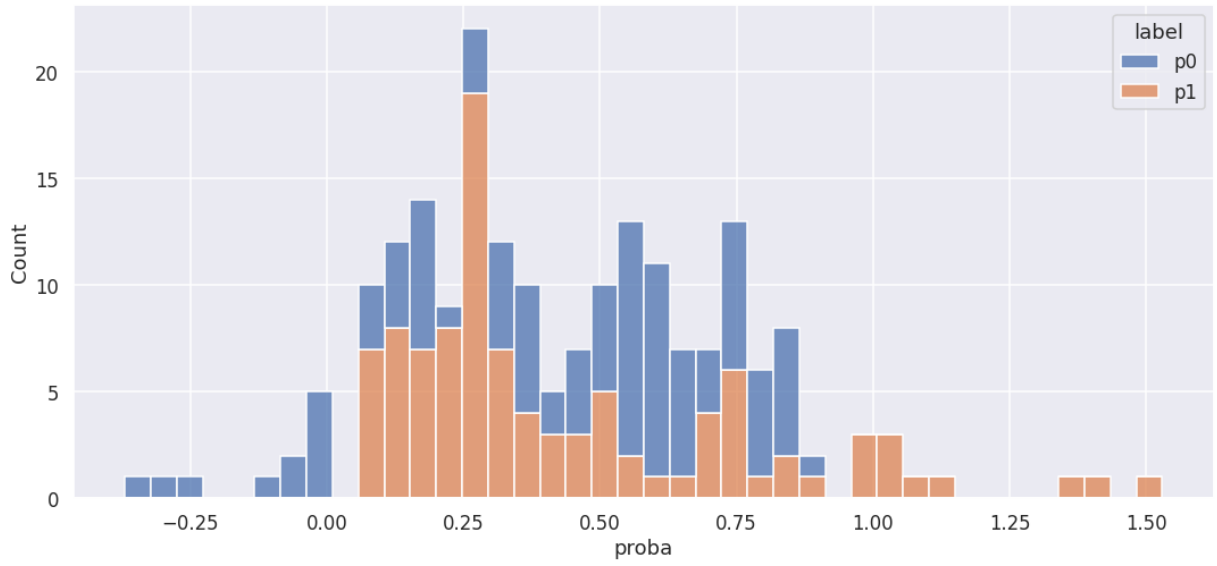

Figure4: CSS results for last layer activation of GPT-J model (pretrained model from transformers python library) for the test set of the dataset ETHICS-commonsense-ambiguous. Results can be found in the github repository.

References

1. Burns, Collin and Ye, Haotian and Klein, Dan and Steinhardt, Jacob. (2022) *Discovering Latent Knowledge in Language Models Without Supervision*. ArXiV.
2. Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song and Jacob Steinhardt. (2021) *Aligning AI With Shared Human Values.* Proceedings of the International Conference on Learning Representations (ICLR)
3. Dataset Amazon Polarity, https://huggingface.co/datasets/amazon_polarity/viewer/amazon_polarity/test