Limitations

Our study focuses solely on the `bert-large-cased` model, meaning our findings cannot be directly generalized to other Pretrained Language Models (PLMs). Future work should extend this analysis to a broader range of architectures to assess whether our observations hold across different models.

Additionally, both our dataset and model were in English, limiting our conclusions to this linguistic context. Since negation varies across languages in both syntax and semantics, evaluating models trained on other languages would be necessary to determine the broader applicability of our approach.

A Attemps to define the reciprocal of the Negator

Instead of learning both a Negator and Affirmator, we also tried to learn a Negator, and then define its reciprocal, to serve as Affirmator. This supposes to define the reciprocal of activation functions and of linear combinations.

To this end, we used bijective activation functions, whose reciprocal functions are:

$$LeakyRELU^{-1}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \frac{x}{\alpha}, & \text{otherwise} \end{cases}$$

$$ELU^{-1}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ log(\frac{x}{\alpha+1}), & \text{otherwise} \end{cases}$$

The reciprocal of the linear combination with parameters W and b, requires W to be invertible (which is why we chose square parameter matrices), and is written as:

$$x = W^{-1}(y - b) \quad (1)$$

Unfortunately, we empirically observed across various runs that the resulting Negator contained at least one non-invertible matrix (namely with a rank lower than the shape of the matrix).

We also tried to use Moore-Penrose pseudoinverse parameter matrices[9]. In such a case, the definition of the reciprocal is as below (with W^{PI} the pseudoinverse of W):

$$x = W^{PI}(y - b) \quad (2)$$

So for a given linear layer, if $y = Wx + b$ then we compute a x' such that $x' = W^{PI}(y - b) \simeq x$, namely there exists a matrix M such that $x' = x + M$.

This empirically failed in the sense that when applying the reciprocal functions in sequence, we noted the M matrices kept growing exponentially. The approximation made using pseudoinverses led to growing errors.

We conclude to the impossibility of inversing the Negator to obtain an Affirmator.

B Reconstruction and Preprocessing of Negated LAMA

The original LAMA dataset is available both in the repository of Petroni et al. (2019) [10] and on the Hugging Face (Wolf et al., 2020)platform [11]. However, with the exception of the SQUAD subset, the number of entries differs between these two sources for every subset. A comparison of the dataset sizes from these two sources can be found in Table 8, under the columns "LAMA" (repository from the original paper) and "LAMA HF" (Hugging Face platform).

The inputs of negated LAMA are either explicitly provided, or through the introduction of a negation pattern.

Upon examining the data, we found that not all entries could be used to reconstruct the negated LAMA dataset. We applied filtering criteria to exclude entries with the following issues:

- Presence of multiple masked tokens
- Absence of a corresponding negated sentence
- Lack of alignment with a recognizable negation pattern

These inconsistencies accounted for nearly two-thirds of the data in the Google-RE and T-REx subsets, and we were unable to fully resolve all of them.

The final sizes of the subsets used to evaluate our models are listed in the "Retained Examples" column of Table 8.

Consequently, we use a version of negated LAMA that is different from the one used by Kassner and Schütze (2020) and Hosseini et al. (2021).

[9] Using pytorch, `https://pytorch.org/docs/stable/generated/torch.linalg.pinv.html`, Paszke et al. (2019))

[10] `https://dl.fbaipublicfiles.com/LAMA/negated_data.tar.gz`

[11] `https://huggingface.co/datasets/facebook/lama`

Dataset	LAMA Subsets		
	LAMA	LAMA HF	#retained examples
SQUAD	305	305	301
conceptnet	2996	29774	8296
Google-re	5527	6106	2926
T-rex	34039	1304391	16991

Table 8: Subset sizes of LAMA from different sources. **Col. "LAMA"**: Number of entries in Petroni et al. (2019) repository. **Col. "LAMA HF"**: Number of entries in the Hugging Face version. **Col. "Retained Examples"**: Final number of entries used in our negated LAMA version.