

Rethinking Relation-Specific Neurons in Language Models

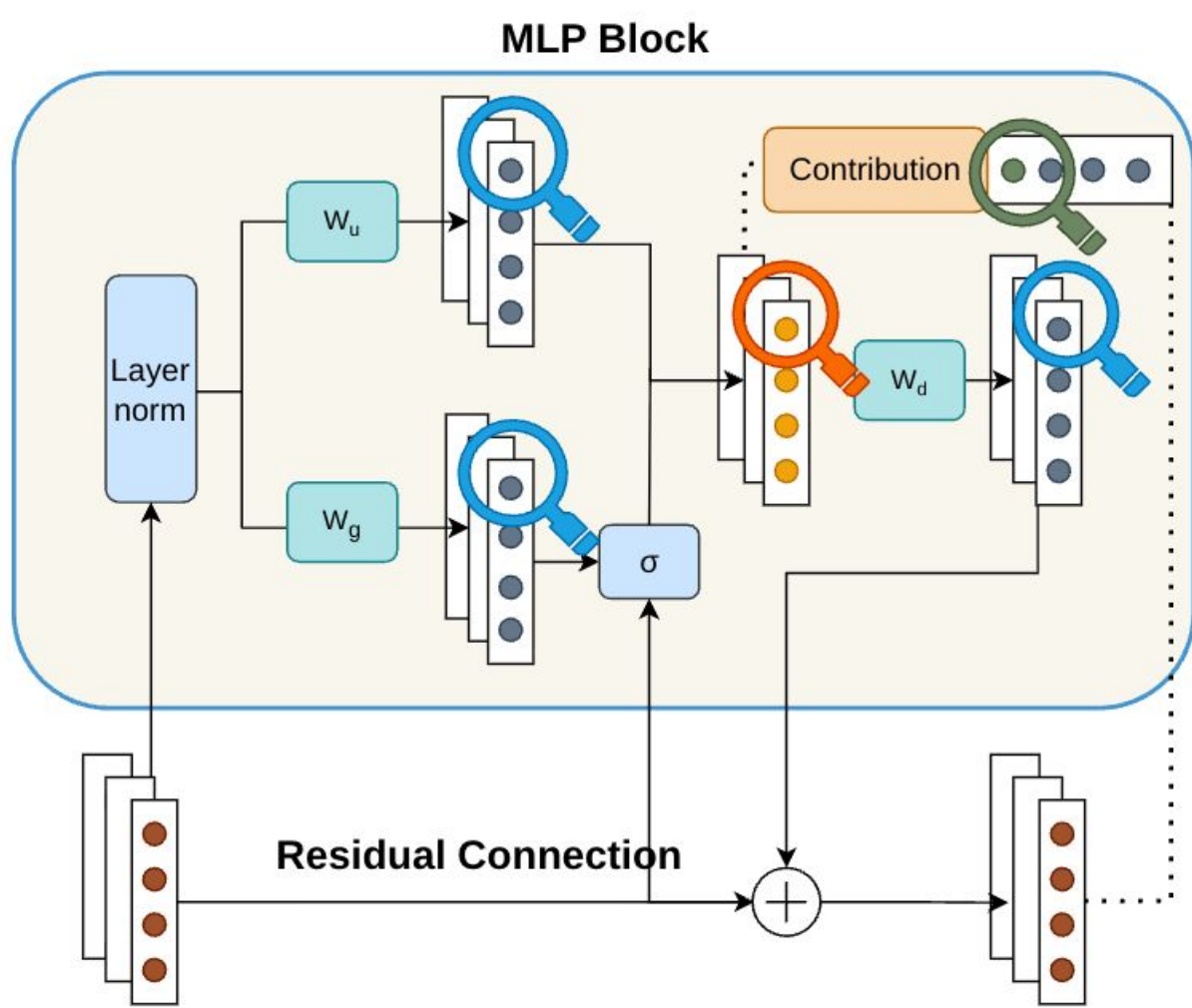
Lea Hirlimann, Sebastian Gerstner, François Yvon, Hinrich Schütze
Mechanistic Interpretability Workshop, ICML 2026



12
Relations

5 MLP
Signals

3 Models



- ColVec** activations of up-, gate- and down-projections
 - FullNeu** activations **after** the activation function
 - Contrib** neuron contributions to the residual stream
- average & last token position

- calculate **Average Precision** scores
- rank neurons based on each Signal
- + Combined Ranking of Contrib & ColVec**
- test ablation effect of top 100-17783 neurons

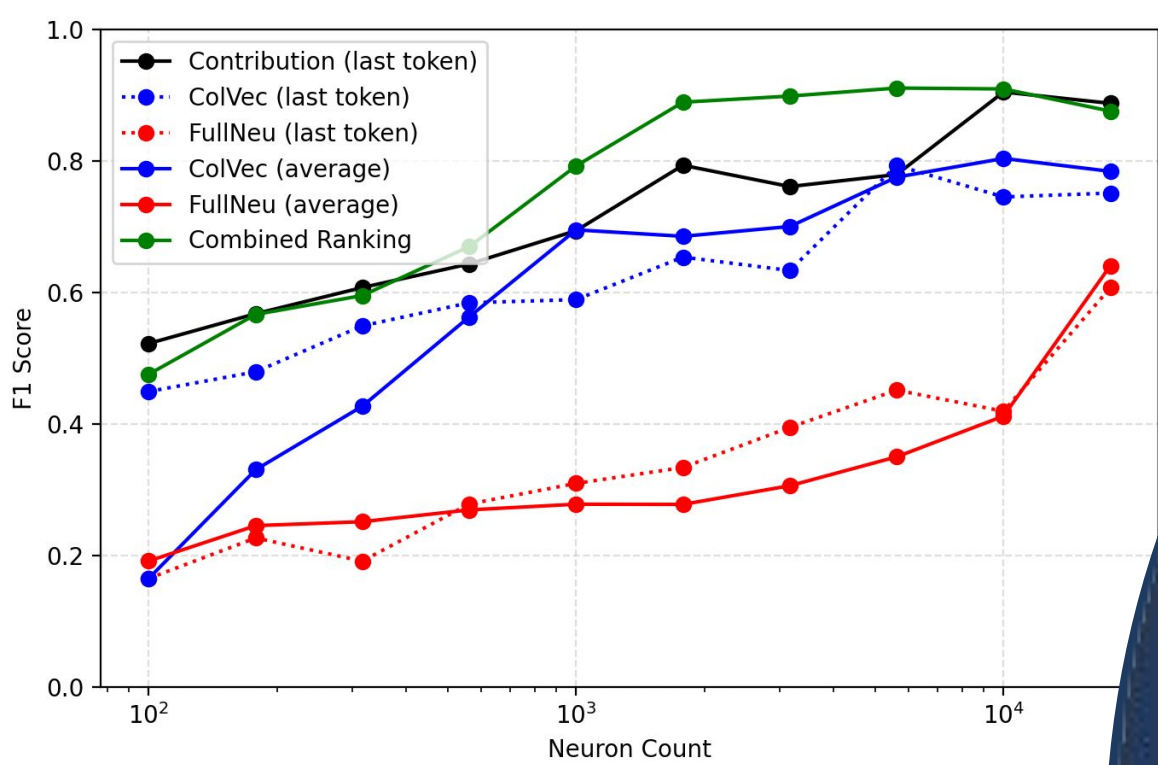
Neuron Ranking & Ablation

Custom F1 metric

- Recall**: % of prompts the model loses the correct answer due to ablation
 - Precision**: average % of preserved answers in all other relations
- strong **target effect**
→ penalize **collateral damage**

1 The signal is your choice.

- ColVec** and **Contrib** return **different sets of neurons** (<10% overlap)
→ **both neuron sets** show good ablation success!
- The **Combined** ranking exceeds both methods individually
→ better average ablation success across all models
→ smaller set of relation-specific neurons
- example: **company_headquarters** training ablation success across neuron counts from 100 to 17783 neurons



2 One size fits none.

- Representations of different relations **vary** in the **number of neurons** involved
- We achieve **100% ablation success** with ...

17783
neurons

for person_plays_instrument

or

1778
neurons

for person_sport_position

316
neurons

or

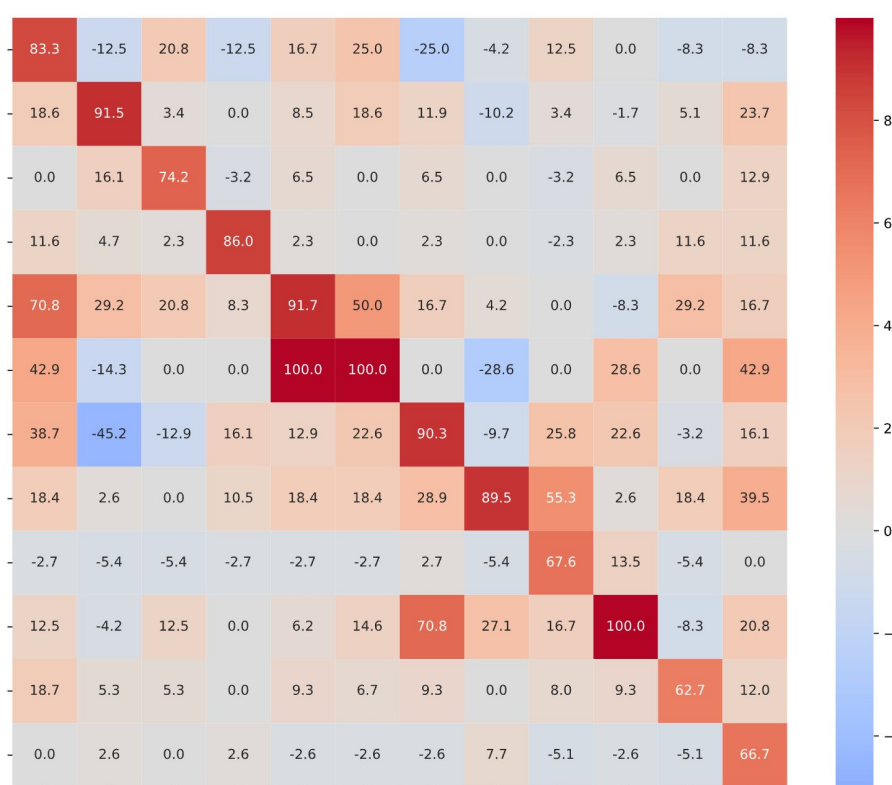
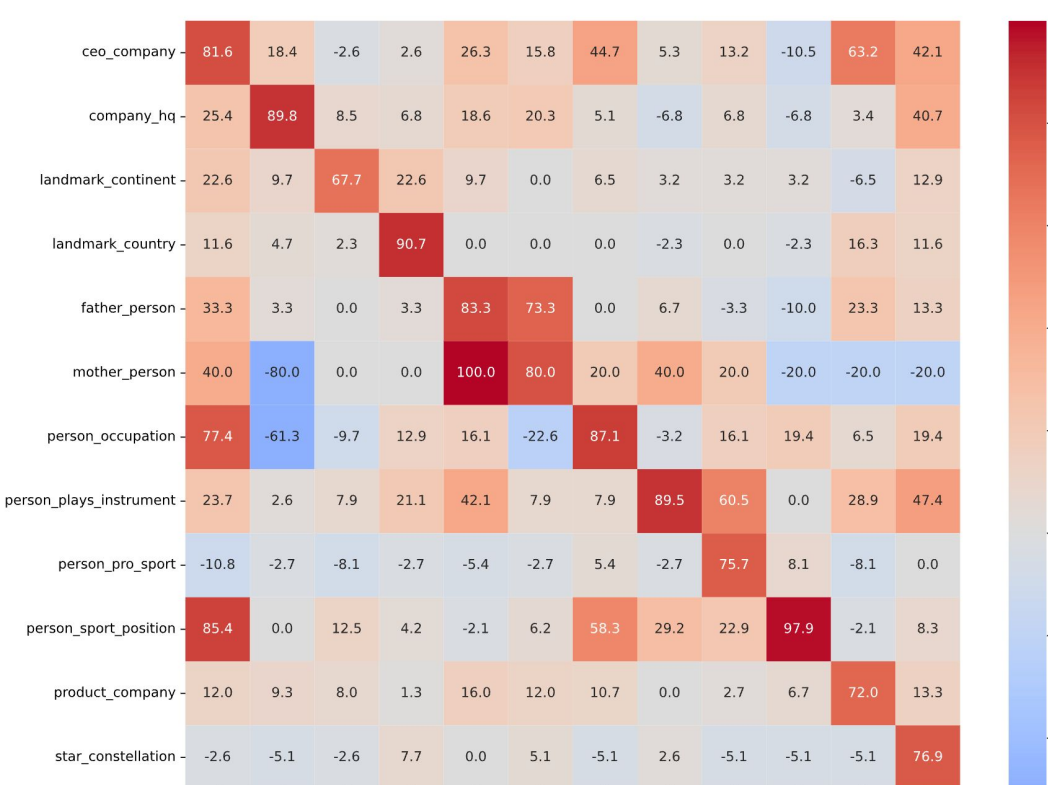
for star_constellation

- Considerably **more or less neurons** fail to capture the relations representation

- Some subjects appear in **multiple relations**
- reversing relations** such as person_mother to mother_person removes that subject overlap
- We show that **effects on other relations** than the target relation is **structural entanglement of representations** and not an artifact of subject overlap

Without subject overlap

With subject overlap



Choose your method.
Choose your answer.

- The categories of the dataset do not necessarily map to **one uniform representation**

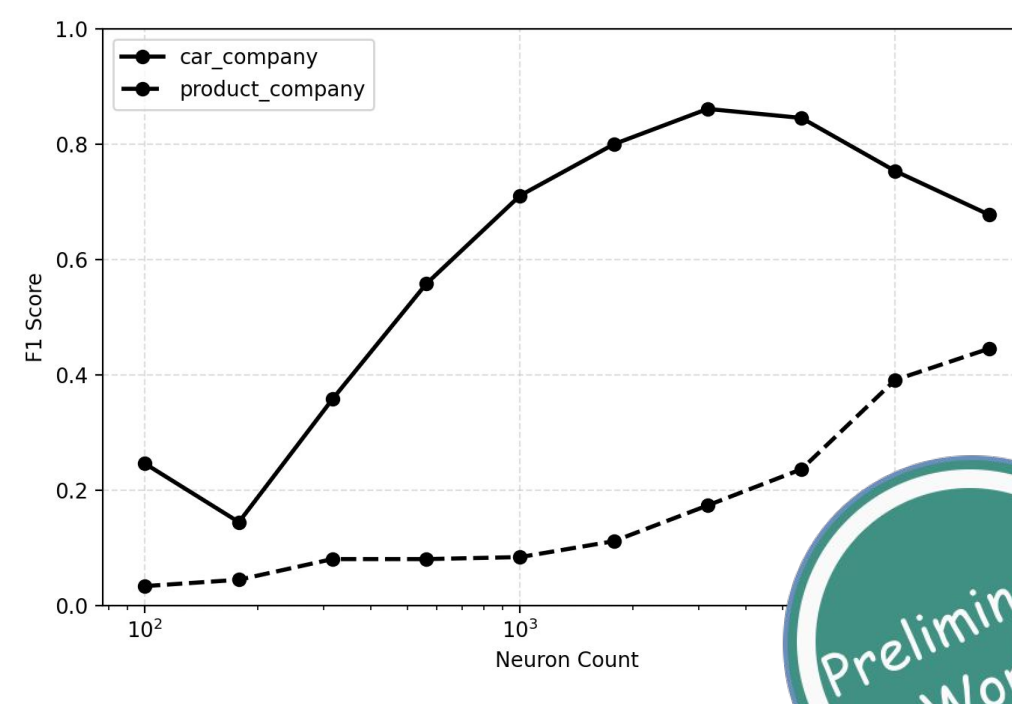
- We **replace** the low success relation **product_company** with its subcluster **car_company**

- This new relation is selected through manual investigation of t-SNE clusters and **fits the internal structure** of the model resulting in a big **improvement** of ablation success

t-SNE activation clusters



Performance comparison



Preliminary Work

3 ... no accidents.

4 Error: Human category not found.



hirlimann@cis.lmu.de



Munich Center for Machine Learning